Beyond Traditional Classifiers: Confidence-Informed Models for PPG Peak Classification

Pivetta Federico
Person code: 10932991
federico1.pivetta@mail.polimi.it

Naclerio Andrea
Person code: 10934883
andrea.naclerio@mail.polimi.it

***Abstract*—This study proposes a methodology for the classification of Photoplethysmography (PPG) signals, crucial for monitoring cardiovascular health and identifying abnormal patterns such as Premature Atrial Contractions (PACs) and Premature Ventricular Contractions (PVCs). Two models based on machine learning and deep learning, addressing binary and multi-class classification, are developed to classify PPG signals into pathological and non-pathological classes. The multi-class model further classifies into Normal (N), premature supraventricular (S), and premature ventricular (V) classes. The study aims to provide the classification for each peak and a measure of the confidence of the model. The dataset comprises 105 PPG signals recorded from 105 patients, annotated with systolic peak positions and ground truth classes. Preprocessing includes resampling, windowing, and filtering to prepare the data for model implementation. Exploratory data analysis involves feature extraction and statistical analysis. The binary classification utilizes an ensemble approach, emphasizing recall for pathological class identification. For multi-class classification, two consecutive binary models are employed, with the second model classifying S and V classes. Although the multi-class algorithm is limited, the binary model shows promising results. Finally, challenges, limitations and potential applications are discussed and further improvements and considerations for real-time application are suggested for future work.**

***Index Terms*—PPG, photoplethysmogram, deep learning, machine learning, LSTM**

I. Context

The goal of this project is to build two models that classify photoplethysmogram, PPG, and peaks. The models are aimed to solve binary and multi-class classification problems and they should provide an estimation of the confidence of the prediction. The multi-class model classifies into Normal (N) premature Supraventricular (S) and premature Vetricular (V) classes while the binary classifier classifies an input sample into pathological (S and V) and non-pathological (N). Moreover, one metric should be defined to monitor and evaluate the performance of the models.

II. Introduction

Photoplethysmography (PPG) signals record the variations in blood volume in the microvascular bed of tissues. PPG signals are commonly obtained from peripheral locations, such as the fingertips or earlobes. These signals are valuable for assessing cardiovascular health, monitoring blood oxygen saturation, and detecting various physiological conditions. Similar to ECG signals [3]–[6], PPG signals can be analyzed to identify abnormal patterns or events as Premature Atrial Contractions (PACs) and Premature Ventricular Contractions (PVCs). The presence of frequent PACs and PVCs is associated with an elevated risk of an unfavorable prognosis. So a monitoring and accurate recognition of these premature contractions are crucial for identifying potential underlying cardiovascular conditions. Therefore, it is crucial to classify correctly all the patients that are potentially affected also allowing to include healthy patients who will then be filtered with successive and more detailed analyses.

A Dataset

The dataset consisted of 105 PPG signals recorded for 30 minutes from 105 patients with 2 different sampling frequencies; specifically, 62 patients were recorded with a sampling frequency f_s = *128 Hz*, and the remaining 43 with f_s = *250 Hz*. Moreover, the signals were annotated with the positions of systolic peaks and their corresponding ground truth classes.

III. Material and methods

A Preprocessing

A preprocessing phase was applied to the raw data in order to make them more suitable for the implemented models.

A.1 Resampling

First, the sampling frequencies were made homogeneous in the whole dataset, choosing *128 Hz* as the standard one. All the signals with a sampling frequency of *250 Hz* experienced an undersampling process. This procedure ensured that signals had the same number of samples in a specific interval and so prevented problems in adopting a model with fixed input dimension. Moreover, the peak positions in the re-sampled signals were updated, since the number of samples changed.

A.2 Signal Windowing

The final task was to classify each single PPG peak separately. In order to deal with this requirement, it was necessary to split the whole signal in windows containing single peaks so the model would be able to classify each of them. To choose the proper window length, correlation between consecutive peak labels was considered, so finding any relevant relation

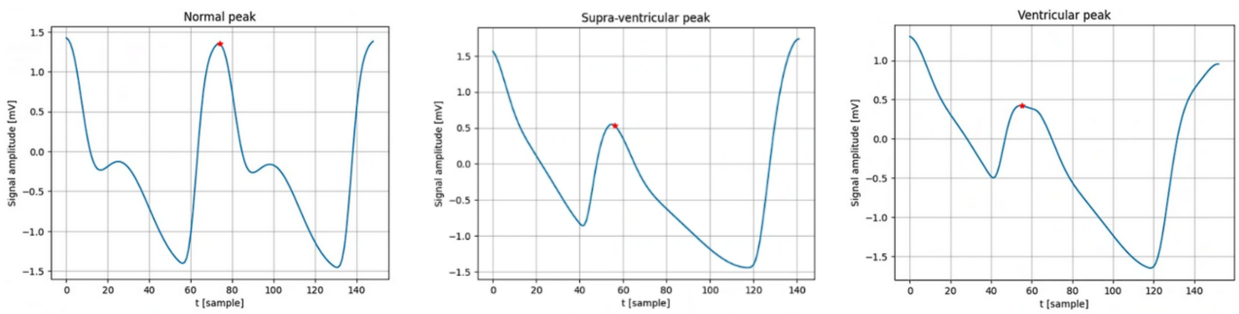

Fig. 1. Example of peaks.

between the current peak and the neighbouring peaks class. Several intervals were tested, with a maximum of 5 previous and 5 after peaks, but no evident correlations were found. For this reason, the window length was chosen in order to extract each single peak individually. A naive approach could be to use a fixed window length. The problem of using this method is tuning a proper length; there could be too large peak intervals where the selected window is cutting the peak, losing relevant information. So, each single peak was extracted considering a window that spanned between the previous peak and the next peak. In this way, it ensured that all the information regarding the complete cardiac cycle was present inside the window. With this approach, both the first and the last peaks of each signal were discarded. Since 93.33% of them belonged to the healthy class, most of the pathological samples were preserved.

A.3 Filtering

Once all the single peaks were extracted, they underwent a filtering process in order to detect and discard possible signal outliers due to instrumentation malfunctioning during the recording, but at the same time denoising corrupted signals.

A.3.1 Length FilteringA first filtering was computed considering the window lengths. A minimum heartbeat of 40 bpm and a maximum one of 180 bpm were considered in order to define a physiological window length range that was used in order to filter the data. Since each peak was extracted considering a window that went from the previous till the next peak, the overall length was not affected by the presence of premature beats. In particular, it could be seen that the premature beats were surrounded by physiological beats. So if the central peak was premature, the distance with the previous peak was shorter but the one with the following was larger; so with this type of extraction, the window lengths would not be affected by the presence of premature beats. In this way, this filter procedure would discard windows completely at random since they were not correlated with respect their class.

A.3.2 Amplitude Filtering The ideal signal that should be recorded was a very smooth signal with an amplitude bounded between -2mV and +2mV, see fig.1. During a real recording, artefacts and noise could change completely the range and the morphology of the true signal. For this reason, a second filtering process was applied considering the amplitude of the signal. In order to introduce some margin, an amplitude range between -4mV and +4mV was considered to not reduce drastically the dataset dimension but still keep only the signal with the characteristic morphology. Looking through the filtered data, signals with a higher amplitude range with respect to the standard one were affected by a high-frequency noise. For this reason, windows with a maximum amplitude between +2.5mV and +4mV or minimum amplitude between -4mV and -2.5mV were extracted and a 2^{nd} order butter low pass filter at 10Hz was applied. If the filter was applied directly to the raw window, an artefact was introduced in correspondence with the initial part of the signal changing a relevant component on the signal morphology. In order to prevent it, initial padding was added to the raw data before the filtering, so the artefact was not affecting the true signal, and after that it was discarded, restoring the initial signal as shown in fig. 2. The adopted padding consisted of 30 samples tuned according to the initial value of the windowed signal. This atypical procedure was used in order to try to preserve in all the way the morphology of the signal; this was crucial since in the case of deep learning approach, models worked directly on the raw data where the shape was the only available information.

A.4 Length Standardization

The adopted splitting strategy kept all the information regarding the current peak, but the different extracted windows had an incoherent length that was strictly related to the patient's heartbeat. For this reason, additional padding was necessary in order to make all the windows equally long. In particular, the same amount of padding was added at the end and beginning of the window in order to not alterate the relative position of the peaks.

B Exploratory Data Analysis

In order to identify differences between signals belonging to different classes (e.g. S vs. V), different features were extracted. In particular, traditional statistics were extracted from the windowed signals: mean, standard deviation, and root mean square. Additionally, kurtosis, entropy, and skewness were extracted as suggested in [1]. Moreover, to provide

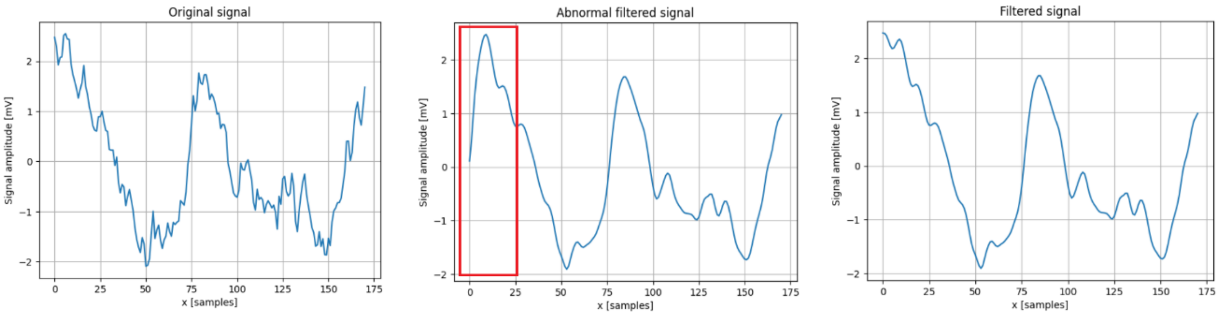

Fig. 2. Amplitude filtering example.

specific information about the position and the amplitude of the peak two parameters were introduced:

$$\Delta T(i) = [T(i) - T(i-1), T(i+1) - T(i)] \qquad (1)$$

$$\Delta A(i) = [A(i) - A(i-1), A(i+1) - A(i)] \qquad (2)$$

The metric of ΔT was *samples*, while the metric of ΔA was mV. All the metrics were computed for each i peak. The class was the only dividing criterion, indeed subjects were merged to prevent selection bias. Except for ΔT, there was no need to normalize because all the other statistics were distributed in about [-1,1]. For what concerns the difference in time, min-max normalization was performed. Scaling parameters were obtained from the training set and used in the training, validation and test sets. These statistics easily discriminated class N from classes S and V. For example in fig.3 the histogram of the standard deviation of class N is very distinct concerning the others that are similar to each other. The inter-class similarity is visually depicted in fig.4. Since no discriminatory patterns of classes S and V were identified, other metrics were attempted. The idea was to identify the presence of a correlation of difference in time and amplitude between the previous peak and the next one. Although different combinations of ΔT and ΔA were implemented, none provided significant results, therefore no additional statistic was considered.

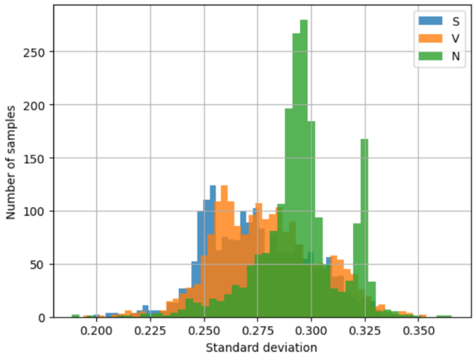

Fig. 3. Histogram of the standard deviation of the signals.

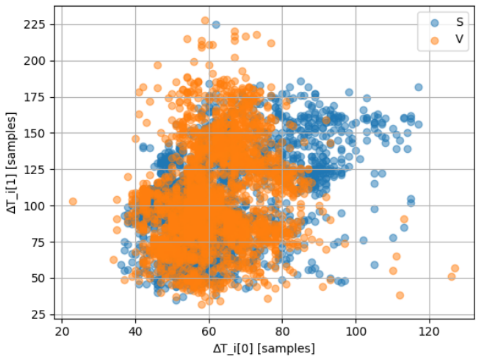

Fig. 4. Scatterplot ΔT.

C Data splitting

Data were splitted into training and validation sets. Test set was avoided because an external test set will be provided at the end of the assignment. To prevent bias, data-splitting was performed by keeping patients divided between training and validation. Moreover, the ratio of samples belonging to classes S and V was kept as much similar as possible across the classes. The splitting ratio was fixed equal to 0.8 and 0.2 respectively. The distributions of samples, divided per class, in the training and validation are shown in fig.5.

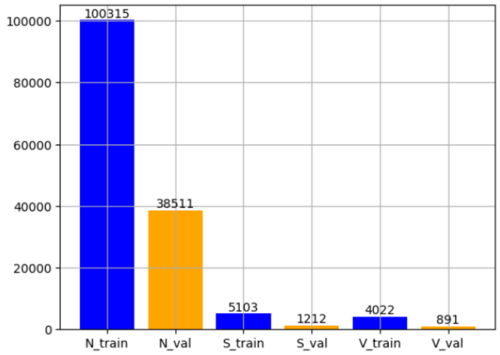

Fig. 5. Amplitude filtering example.

D Models

Different methodologies were adopted to address the binary and the multi-class classification problems. For what concerns the binary classification, to solve the problem of unbalanced data, an ensemble approach was used. Once identified the most performing model, it was trained 5 independent times using 5 different training sets obtained by merging the same group of pathological samples (S, V classes) and 1/5 of the non-pathological samples (N class) respectively. A dense layer with one single neuron and *sigmoid* activation function was used as last layer of the network. Finally, the post-processing consisted in identifying a threshold to classify into class 0 (non-pathological) and class 1 (pathological) though the following formula:

$$y = \begin{cases} 0 & \text{if } out(i) \leq \tau \\ 1 & \text{if } out(i) > \tau \end{cases} \tag{3}$$

where i, $out(i)$, $y(i)$ are respectively the input sample, the output of the network and the post-processed prediction. The model was aimed at reducing false negatives, FN, due to its clinical purpose. Nevertheless, its performance should be kept acceptable for both classes. Therefore, the loss function was weighted in order to increase the importance of pathological samples and the value of τ was fine-tuned. Different metrics were adopted to monitor the training of the models, the most important one was the recall:

$$recall = \frac{TP}{TP + FN} \tag{4}$$

The reduction of FN is fundamental to prevent positive patients not correctly classified. The selection of the best model was performed considering this metrics, it was monitored also in the rescheduling of the learning rate while binary cross-entropy on the validation set was preferred in the early stopping in order to prevent models that were too unbalanced to the pathological class. For what concerns the multi-class classification task, two consecutive binary models were implemented as shown in fig. 9. This choice was made to improve the performance of the algorithm, to prevent the issue of unbalanced data and to increase the control of the behaviour of the model. More in detail, the model adopted in the binary task was used to classify between non pathological samples and pathological samples. The last were used as input of the second binary model, called *SV-classifier*, that classified into S and V classes. Different methodologies were used to address the second binary problem. In particular, using signal statistics, see subsection III. B, different machine learning models were implemented. In detail, Random Forest (*RF*), Support Vector Machine (*SVM*), and K-Nearest Neighbors (*KNN*) were trained. Their hyperparameters were optimized using GridSearch on a 3-fold cross-validation. Moreover, Principal Component Analysis, *PCA*, was performed in order to reduce the dimensionality of the input feature space, allowing to obtain simpler models.

The task was approached using raw data and deep learning models. Several architectures were trained to solve SV-classification. For what concerns 1D-convolutional networks, *VGG16*, ResNet-style architecture (*RN*) and ResNet-style architecture with squeeze and expand blocks, *RN-SE*, were adopted. Furthermore conv-1D architecture with multiple recurrent layers, in particular Long-Short-Term-Memory, *LSTM*, and Bidirectional-Long-Short-Term-Memory, *BLSTM*, were used. All the models were optimized and their accuracy, precision and recall on the validation set were compared. Finally, the best model was enhanced by including as input data the features of the samples, *FEAT*, and/or its Fast Fourier Transform, *FFT*, [2]. FFT information was included to help the model to identify oscillations that could be discriminatory for one specific class.

E Confidence computation

Class predictions were combined with the estimations of the confidence. For what concerns the binary classification, it was computed starting from the output of the neuron in the last layer characterized by a sigmoid activation function. In particular, if the prediction is class 1, $y(i) = 1$, the confidence, $c(i)$, is equal to the output $out(i)$. While if the prediction is class 0, $y(i) = 0$, the confidence, $c(i)$, is proportional to $1 - out(i)$. The proportion factor depend on the value of τ and it is computed by mapping the interval $0 - \tau$ in $0 - 0.5$ in order to have a consistent result. Since τ was fixed $\tau > 0.5$, the value of the confidence of samples predicted as pathological was greater than 0.5. The choice of not scaling this result was to not decrease its value and therefore to maintain the importance of pathological predictions. For what concerns *SV-classificator*, the last layer was a dense layer with 2 neurons and softmax activation function. Therefore the confidence of the prediction of the model was estimated directly from the output itself. Regarding the multi-class classification problem, the confidence is computed differently for the three classes. The confidence of the non-pathological class is directly provided from the first classifier, while the confidence of the pathological classes is the result of the multiplication of the confidences of the first and second models. Finally, during the inference phase, samples at the beginning or at the end of the input signal and samples that are discharged due to the filtering process are automatically predicted to class N because it's the most numerous class. To take into account pathological peaks that are misclassified as physiological ones, the confidence of these predictions is assigned equal to 0.5 in the binary classification and equal to 0.33 in the multi-class problem.

IV. Results

As described in section D, the binary model was an ensemble model and it merged the prediction of 5 models. In fig. 10 is shown the loss function curve during the training process of one of these models. The performances of the ensemble model on the whole validation set are reported in table I. They were reported both classes-divided and classes-merged to visualize the importance of recall in the pathological class. Finally, the

Receiver Operating Characteristic, *ROC*, curve is depicted in fig. 6. It was obtained by varying the value of τ in formula 3.

	Precision	Recall	Accuracy
Non-pathological (0)	0.99	0.95	-
Pathological (1)	0.50	0.91	-
Model	0.75	0.93	0.95

TABLE I
PERFORMANCE METRICS OF THE BINARY CLASSIFIER.

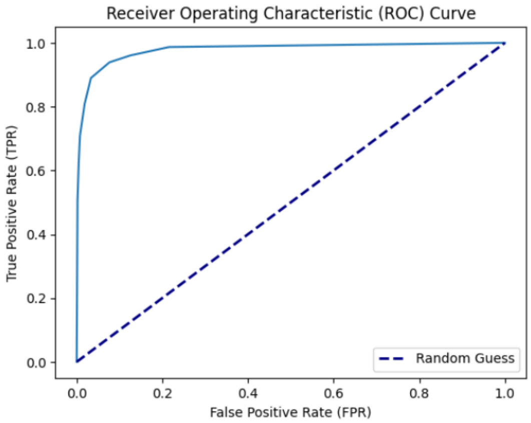

Fig. 6. ROC curve of the binary model.

For what concerns the multi-class problem, it was approached through a double binary classification as discussed in section D. *SV-classifier* was aimed to classify only pathological samples into classes S and V. The performances of the implemented models on the validation set are shown in table II. The best models, one for each approach category, are bold. *BLSTM* resulted the best model, therefore it was improved adding samples' features, *FEAT*, and frequency information, *FFT*. Their results on the validation set are shown in table III. Finally, the performance of the multi-class model, obtained by combining the binary model and *SV-classifier*, computed on the validation set are shown in table IV and its confusion matrix is shown in fig.7. The confusion matrix is evaluated on a subset of the validation set, in particular, samples belonging to the N class were undersampled to balance the classes and provide a more interpretable result.

Approach	Model	Precision	Recall	Accuracy
Machine Learning	SVM	0.47	0.48	0.44
	RF	0.56	0.56	0.56
	KNN	**0.56**	**0.57**	**0.57**
	KNN-PCA	0.52	0.52	0.52
Deep Learning	VGG16	0.55	0.54	0.53
	RN	0.57	0.57	0.55
	RN-SE	0.51	0.50	0.45
	LSTM	0.58	0.58	0.57
	BLSTM	**0.61**	**0.61**	**0.62**

TABLE II
PERFORMANCE METRICS OF *SV-classifiers*.

Model	Precision	Recall	Accuracy
BLSTM-FFT	0.59	0.59	0.59
BLSTM-FEAT	0.56	0.56	0.55
BLSTM-FFT-FEAT	**0.62**	**0.62**	**0.63**

TABLE III
PERFORMANCE METRICS OF IMPROVED *SV-classifiers*.

V. Discussion

From table II it can be seen that the Bidirectional-LSTM (BLSTM) resulted the most performing model. The first 2 blocks were made of a bidirectional LSTM, followed by conv-1D and max-pooling layer in order to extract relevant features considering the whole shape of the input signal. Other 2 conv-1D and max-pooling layers were added to increase the receptive field. Follows a global average pooling layer (GAP) that converted the output volume in a 1D vector that was used as input to a final dense layer to predict the class. Since the model needed to manage sequential data, where the input order is crucial, the BLSTM had the most adapt structure with its internal hidden state that, acting as a memory, was able to extract relevant features by looking at the correlation between the inputs.

Thanks to its performances, the same structure was used also for the initial binary classifier to distinguish between pathological and non-pathological samples. The results, in this case, were reasonably good, since the signal shapes were significantly different between each other and so the model was able to extract relevant features, since they are extracted directly from the raw data, that allow to distinguish between the 2 classes On the contrary, the *SV-classifier* had poor performance since the input data that were used, V and S windows, have very similar morphologies thus the features extracted by the model were not distinctive; so the model had the tendency of confusing the two classes as can be seen from the performances shown in table III. But looking at the accuracy reported in table IV, it could be thought that the general performance of the model in distinguishing between

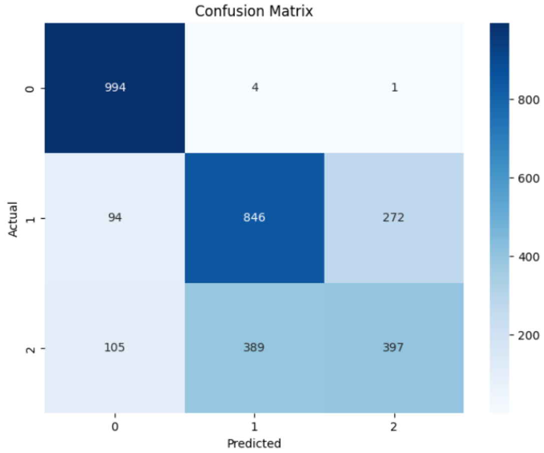

Fig. 7. Confusion matrix of the multi-class model.

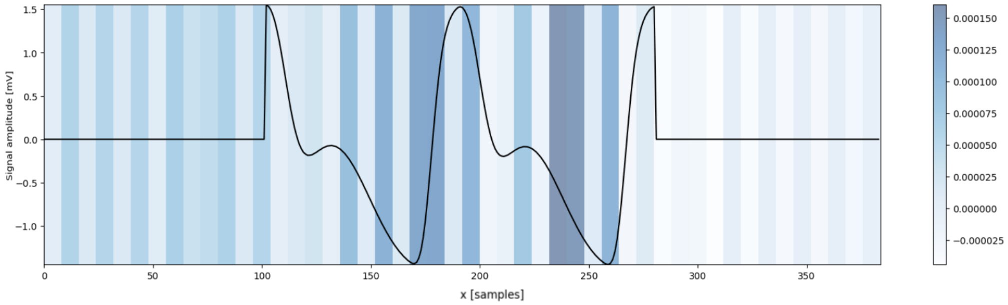

Fig. 8. Grand-CAM on an example signal.

the 3 classes, N, S and V, is considerably good. But it is important to underline that this metric is not relevant if the classes are considerable unbalanced.

To better understand how the model works, a final explainability method was adopted, GradCAM [7]. In particular the GradCAM is a technique able to give an explainability heatmap extracted from the last convolutional in order to enhance which are the most relevant input regions that are used by the model in order to assess the predicted class. As shown in fig.8 the model tends to focus on the ascending and descending region of the central peak, considering both the slope and the peak value. Anyway, this technique gives a coarse explanation of the model since the heatmap is much shorter with respect to the original input length, but it's still relevant since it's a first step to understand what the model is looking for to classify the samples.

As already mentioned in the subsection III.D, the adopted model structure allows us to better manage the performance of the model and minimize the false negatives but, on the contrary, it requires very long inference time. As a consequence, a model implementation for integrated application in devices that acquire PPG signal become unfeasible due to the the resource and time demanding. In order to allow a device implementation, the initial classifier could be used thanks its good performance. Moreover in order to make it more 'light', the model could be trained on a larger balanced dataset to drop the ensemble design.

VI. Conclusion

As discussed in section V, the performance of the binary model is acceptable while *SV-classifier* and therefore the multi-class model is not able to obtain significant results. Consequently, the proposed methodology can not be used as an automatic diagnostic tool for premature supra-ventricular complex and premature ventricular complex identification. Nevertheless, the binary model could be integrated into the PPG acquisition system as a support tool to help clinicians in the screening phase of long acquisition. In particular, the proposed algorithm for the identification of pathological occurrences is relatively sensitive and it could help the specialist to focus attention on the most relevant peaks. On the other hand, the models can be improved using more sophisticated architecture and training them with larger and more heterogeneous data. In detail, instead of using recurrent neural networks, transformers with attention can be implemented. They can be easily parallelized allowing to manage of larger data sizes. Moreover, considering the same number of parameters, the inference time would be reduced and they can be used as real-time classification algorithms. For what concerns data, largest datasets can be exploited to increase both the performance of the models and their generalization capability. Moreover, the availability of more pathological samples could potentially show some discriminatory features. Finally, the only use of PPG signals represents a limitation because differences across the classes are very limited. For sure, the combined use of PPG and ECG signals would provide better results as shown by multiple studies published in the scientific literature.

	Precision	Recall	Accuracy
Non-pathological	0.99	0.95	-
Supra-ventricular	0.33	0.70	-
Ventricular	0.32	0.45	-
Model	0.55	0.70	0.93

TABLE IV
PERFORMANCE METRICS OF MULTI-CLASS CLASSIFIER.

References

[1] Ovadia-Blechman Z, Hauptman Y, Rabin N, Wiezman G, Hoffer O, Gertz SD, Gavish B, Gavish L. Morphological features of the photoplethysmographic signal: a new approach to characterize the microcirculatory response to photobiomodulation. Front Physiol. 2023 Sep 25;14:1175470. doi: 10.3389/fphys.2023.1175470. PMID: 37817983; PMCID: PMC10561251.

[2] Cooley, J. W., Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, 19(90), 297-301. DOI: 10.1090/S0025-5718-1965-0178586-1

[3] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, Arash Gharehbaghi, A review on deep learning methods for ECG arrhythmia classification, Expert Systems with Applications: X, Volume 7, 2020, 100033, ISSN 2590-1885, https://doi.org/10.1016/j.eswax.2020.100033.

[4] Liu Z, Zhou B, Jiang Z, Chen X, Li Y, Tang M, Miao F. Multiclass Arrhythmia Detection and Classification From Photoplethysmography Signals Using a Deep Convolutional Neural Network. J Am Heart Assoc. 2022 Apr 5;11(7):e023555. doi: 10.1161/JAHA.121.023555. Epub 2022 Mar 24. PMID: 35322685; PMCID: PMC9075456.

[5] Sraitih, M.; Jabrane, Y.; Hajjam El Hassani, A. An Automated System for ECG Arrhythmia Detection Using Machine Learning Techniques. J. Clin. Med. 2021, 10, 5450. https://doi.org/10.3390/jcm10225450

[6] Jianyuan Hong, Hua-Jung Li, Chung-chi Yang, Chih-Lu Han, Jui-chien Hsieh, A clinical study on Atrial Fibrillation, Premature Ventricular Contraction, and Premature Atrial Contraction screening based on an ECG deep learning model, Applied Soft Computing, Volume 126, 2022, 109213, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2022.109213.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, ”Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

Appendix A

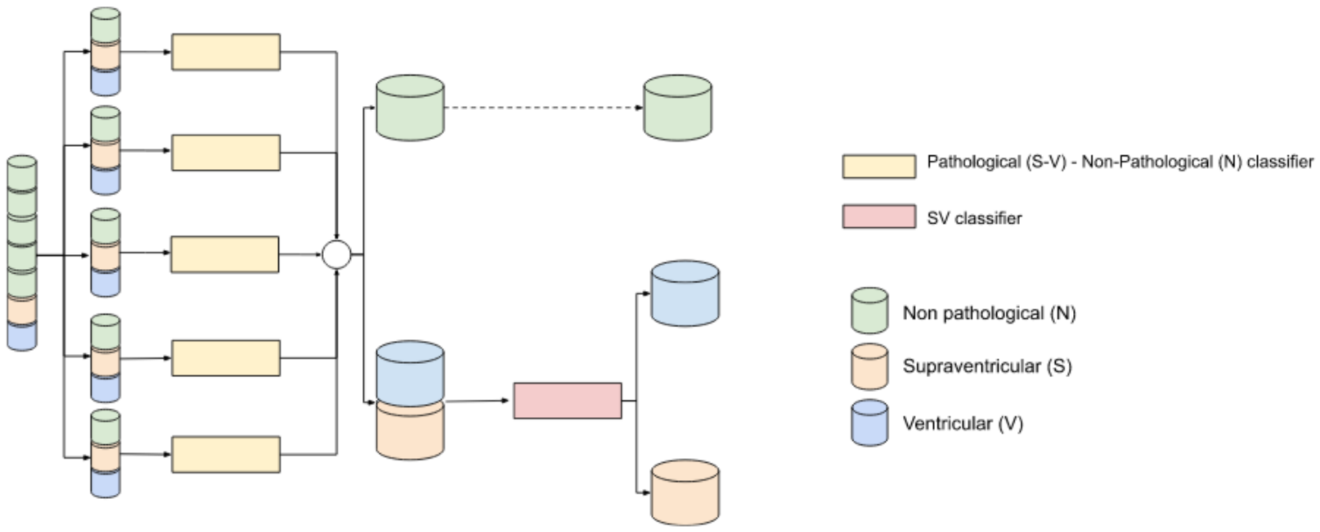

Fig. 9. Multi-class classifier framework.

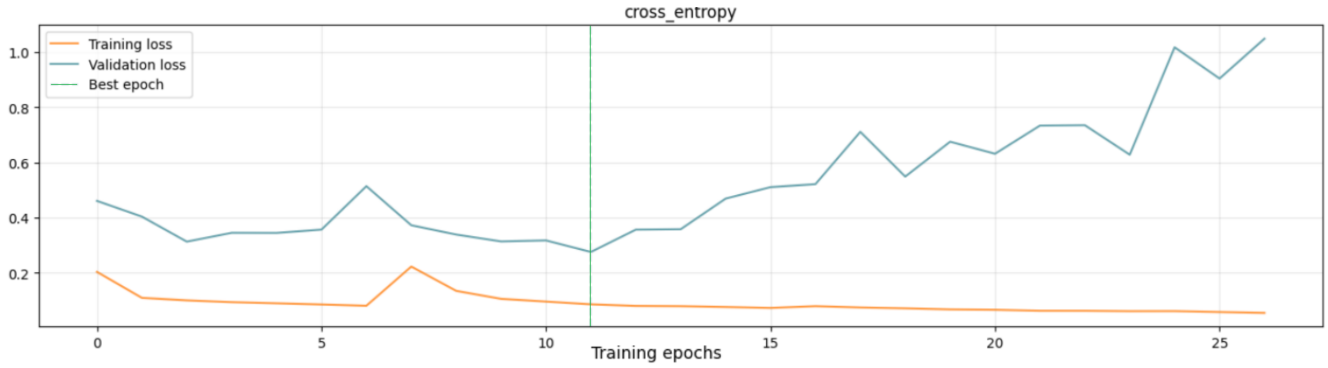

Fig. 10. Trend of the loss function of one binary model.