An hybrid CNN-Transformer model based on multi-feature extraction and attention fusion mechanism for cerebral emboli classification

Yamil Vindas YAMIL.VINDAS@CREATIS.INSA-LYON.FR
CREATIS Laboratory
Univ Lyon, INSA-Lyon,
Université Claude Bernard Lyon 1,
UJM-Saint Etienne, CNRS, Inserm,
CREATIS UMR 5220, U1294,
Lyon, F-69100, France

Blaise Kévin Guépié BLAISE_KEVIN.GUEPIE@UTT.FR
Laboratoire Informatique et Société Numérique
Université de Technologie de Troyes
10004 Troyes, France

Marilys Almar MARILYS.ALMAR@ATYSMEDICAL.COM
Atys Medical
17 Parc Arbora
69510 Soucieu-en-Jarrest, France

Emmanuel Roux EMMANUEL.ROUX@CREATIS.INSA-LYON.FR
CREATIS Laboratory
Univ Lyon, INSA-Lyon,
Université Claude Bernard Lyon 1,
UJM-Saint Etienne, CNRS, Inserm,
CREATIS UMR 5220, U1294,
Lyon, F-69100, France

Philippe Delachartre PHILIPPE.DELACHARTRE@CREATIS.INSA-LYON.FR
CREATIS Laboratory
Univ Lyon, INSA-Lyon,
Université Claude Bernard Lyon 1,
UJM-Saint Etienne, CNRS, Inserm,
CREATIS UMR 5220, U1294,
Lyon, F-69100, France

Editor: Editor's name

Abstract

When dealing with signal processing and deep learning for classification, the choice of inputting whether the raw signal or transforming it into a time-frequency representation (TFR) remains an open question. In this work, we propose a novel CNN-Transformer model based on multi-feature extraction and learnable representation attention weights per class to do classification with raw signals and TFRs. First, we start by extracting a TFR from the raw signal. Then, we train two models to extract intermediate representations from the raw signals and the TFRs. We use a CNN-Transformer model to process the raw signal and a 2D CNN for the TFR. Finally, we train a classifier that combines the outputs of both models (late fusion) using learnable and interpretable attention weights per class. We evaluate our approach on three medical datasets: a cerebral emboli dataset (HITS), and two electrocardiogram datasets, PTB and MIT-BIH, for heartbeat categorization. The results show that our multi-feature fusion approach improves the classification performance with respect to the use of a single feature method or other multi-feature fusion methods. Furthermore, it achieves state-of-the-art results on the HITS and PTB datasets with a classification accuracy of $93,4\%$ and $99,7\%$, respectively. It also achieves excellent performance on the MIT-BIH dataset, with an accuracy of $98,4\%$ and a lighter model than other state-of-the-art methods. What is more, our fusion method provides interpretable attention weights per class indicating the importance of each representation for the final decision of the classifier.

1. Introduction

Signals can be defined as encoded representations of physical phenomena. In the past decade, a lot of works have focused on image classification using deep learning methods such as deep neural networks (DNN) and convolutional neural networks (CNN) (Rawat and Wang, 2017). In comparison, fewer works have focused on signals with a temporal dependence, such as audio signals or sensors signals. Yet, temporal dependency is particularly interesting in the medical field as different devices such as Transcranial Doppler (TCD) ultrasound, electrocardiograms (ECG) or electroencephalogram (EEG), produce signals with a rich temporal dimension.

These time-dependent signals can be used to detect pathologies such as patent foramen ovale (TCD) and arrhythmia (ECG). They can be used to detect potential causes of stroke before it occurs (TCD) by monitoring the cerebral blood flow to detect high intensity transient signals (HITS), which are potential gaseous or solid particles that can circulate in the bloodstream (Wallace et al., 2015). Some works have focused on discriminating artifacts from emboli using signal processing and machine learning techniques (Guepie et al., 2019) and deep learning techniques (Vindas et al., 2022). However, few works focus on portable TCD data (Guepie et al., 2019) and in vivo artifact/gaseous/solid classification (Vindas et al., 2022). This last point is of particular interest as gaseous and solid emboli identification can be useful for transcatheter aortic valve implantation (TAVI) [1] (Aggarwal et al., 2018). However, the classification of solid / gaseous emboli has not been thoroughly studied, as in vivo data acquisition is difficult (Tafsast et al., 2018), and classical signal processing and machine learning techniques have not been able to achieve satisfactory results (Darbellay et al., 2004; Markus and Punter, 2005).

Classical signal processing techniques extract spectral and handcrafted features to classify signals (Purwins et al., 2019). In contrast, deep models automatically extract features from signals or their time-frequency representations (TFRs). To take advantage of the tem-

[1] This procedure generates several gaseous emboli and can generate few solid emboli. It is then important to be able to detect the solid emboli among the numerous gaseous emboli to help clinicians to prevent strokes.

poral context of time-dependent signals, models such as 1D CNNs (Nguyen et al., 2021), recurrent neural networks (RNNs) (Hori et al., 2018), or convolutional deep belief networks (Lee et al., 2009) can be used.

One of the main difficulties when manipulating time-dependent signals with deep learning models is the choice of the optimal representation to use to solve the task. Often, TFRs are used instead of the raw signal (Natarajan et al., 2020; Gong et al., 2021), although the raw signal can provide valuable and complementary information on the studied phenomenon (of the Ninth International Cerebral Hemodynamic Symposium, 1995). Moreover, some works propose to combine different features and / or representations (Chen et al., 2021; Yao et al., 2021), and the optimal way of combining them is a critical issue (Ahmad et al., 2021). In these works, fusion is done by concatenation or majority vote (at different levels) using precomputed representations. None of these works directly uses the raw signal and only Jin et al. (2020) uses attention weights to help interpret the final classification. However, the weight of each representation in the prediction remains unclear.

Inspired from the above-mentioned motivations, we propose an hybrid CNN-Transformer model based on multi-feature extraction and late fusion with learnable and interpretable attention weights. First, we compute the magnitude spectrogram of the raw signal. Then, we feed the raw signal to an hybrid 1D CNN Transformer model and the spectrogram to a 2D CNN model. Two sets of classification predictions are extracted: one focusing on temporal characteristics (from the raw signal) and the other on spectral characteristics (from the spectrogram). Afterwards, these two sets of classification predictions are combined using learnable attention weights per modality and per class. It allows us to interpret the importance of each modality in each class probability predicted by the final model.

Our main contribution can be summarized as follows :

- A novel hybrid CNN Transformer model exploiting both the temporal context thanks to the raw signal and its spectrogram representation.
- We exploit directly the raw signal thanks to an hybrid 1D CNN Transformer model.
- A late-fusion mechanism based on learnable attention weights which are interpretable.
- State-of-the art results on two medical datasets consisting of two different tasks.

The rest of the paper is structured as follows. In Section 2 we introduce some related works. In Section 3 we present the proposed model and its late fusion mechanism in detail. In Section 4 we explain the datasets that we used and how they were pre-processed. In Sections 5 and 6 we provide the experimental setup, and we discuss the results of the different experiments, respectively. Finally, in Section 7 we conclude and give the guidelines to our future work.

Generalizable Insights about Machine Learning in the Context of Healthcare

Several medical devices used for physical examination produce temporal dependent signals as input (TCD, ECG, EEG, etc.). Deep learning approaches (typically CNNs) are often very efficient when working on pre-extracted TFR from these signals but their outputs usually suffer a lack of interpretability. Moreover, few models directly exploit the raw temporal-dependent signal and/or both representations. In this work, we focus on the use of both types of representations (temporal and spectral), as we found that it benefits the model performance on several medical datasets. Finally, we propose a late fusion mechanism based on learnable attention weights, making our final model easily interpretable with respect to each input representation. In summary, our method further pushes the deep model capabilities to exploit time-dependent medical signals while maintaining the predictions interpretable.

2. Related Work

2.1. Multimodal learning

Multimodal learning consists in exploiting complementary representations of a phenomenon to solve a task (Baltrusaitis et al., 2017). In fact, different modalities give complementary points of view that can help improve the performance of a model (Akbari et al., 2021). (Baltrusaitis et al., 2017) establish a taxonomy of the different challenges in multimodal learning. Our work is related to two topics: multimodal representation and multimodal fusion.

Multimodal representations are of two types: joint and coordinated representations. On the one hand, to obtain joint representations, some works start by individually extracting hidden features from each modality, and then they project each representation in a common space (Agrawal et al., 2017; Mei et al., 2016). (Müller, 2007) used autoencoders (AEs) to extract features from each modality and fuse the obtained representations with another AE model. On the other hand, one can coordinate individual representations in order to satisfy some constraint (instead of creating a joint representation). This can be done by minimizing the distance between each representation (Kiela and Bottou, 2014), or structuring the representations through order constraints (Taylor et al., 2012) or correlation (Poria et al., 2015).

Unlike multimodal representation, multimodal fusion is not limited to combining representations from different modalities (Baltrusaitis et al., 2017). Three commonly used fusion techniques are early, intermediate, and late fusion. Early fusion allows correlating low-level features from the available modalities by combining the different modalities before feeding them to the model. (Castellano et al., 2008) combine features extracted from face, body, and speech data before feeding them through a Bayesian classifier (better than single-modality models). Intermediate fusion combines each modality representation before the final decision of the model. (Akbari et al., 2021) used a self-supervised multimodal Transformer to exploit video, audio, and text information on tasks such as video action recognition, audio event classification, and zero-shot retrieval. Late fusion takes different models, each trained with one modality, and combines their outputs. Among the different late fusion approaches we can cite averaging (Rohrbach et al., 2015), weighting (Ouyang et al., 2014), voting (Mckeown et al., 2010), max, or learned combination (Gebru et al., 2018).

In this paper, we focus mainly on joint representations and late fusion of different representations of a single modality.

2.2. Learning with multiple features and representations

Inspired from multimodality and its advantages, several works have focused on ways of combining different representations coming from a single modality.

In computer vision, several papers have tried to combine different features and/or representations of a single modality (Zhu and Jiang, 2020; Tiong et al., 2019). For face recognition (Zhu and Jiang, 2020) used two-dimensional principal component analysis (2DPCA) and local binary patterns (LBP) to extract and combine global and local features from face images. Similarly, (Tiong et al., 2019) extracts different image features (histogram of gradients, LBP, and entropy texture), each passed to a dedicated CNN model and followed by intermediate fusions (concatenation, averaging, and max selection). They feed the obtained fused features to a DNN and combine the outputs using a late decision fusion layer.

In signal processing, different approaches use TFR or other handcrafted features (Jin et al., 2020; Kim and Lee, 2019; Chen et al., 2021; Ahmad et al., 2021). (Jin et al., 2020) did emotion recognition using two successive weighted concatenations of (1) features extracted by an LSTM model from different Mel frequency cepstral coefficients (MFCC) and (2) features extracted by a DNN model from behavioral data. (Kim and Lee, 2019) used a concatenation of three TFRs (spectrogram, mel-spectrogram, and MFCC) with an LSTM to classify power signals. (Chen et al., 2021) use late feature fusion to classify ECG heartbeat signals to detect atrial fibrillation. They compute two features (eigenvalues of the recursive matrix, and coherence spectrum characteristic), that pass through a 1D CNN before a majority voting to combine the models' outputs. (Ahmad et al., 2021) did heartbeat categorization using ECG signals. The authors first extracted three images from the raw signal: gramian angular field, recurrent plot (RP), and Markov transition field. They achieve state-of-the-art performances on two heartbeat categorization datasets (PTB and MIT-BIH) with a multimodal feature fusion (MFF) where each feature is passed to an AlexNet model and then fused before feeding them to an SVM classifier.

3. Methods

In this paper, we propose a novel classification model using temporal-dependent signals and TFRs. The model is composed of two encoder modules (one for the raw signal and one for the TFR) and one classification model with learnable attention weights per modality and per class. Figure 1 shows an overview of our proposed method with two main branches for the extraction of specific features and an interpretable fusion layer.

3.1. Hybrid 1D CNN Transformer encoder

Let's denote the raw signal by $\mathbf{R} = [\mathbf{R_1}, ..., \mathbf{R_N}] \in \mathbb{R}^{N \times C}$, where N is the length of the input signal and C is the number of channels of the signal.

To extract features from the raw signal, we propose to use an hybrid 1D CNN Transformer architecture. The architecture that we used is strongly inspired by (Natarajan et al., 2020) and is resumed in Figure 2. The first blocks correspond to 1D CNN blocks, allowing to efficiently extract features from the raw signal thanks to overlapping 1D convolution filters. The obtained features form the embeddings that are fed to the Transformer encoder (TE). Indeed, one input embedding of the TE is composed of all the channel components ob-

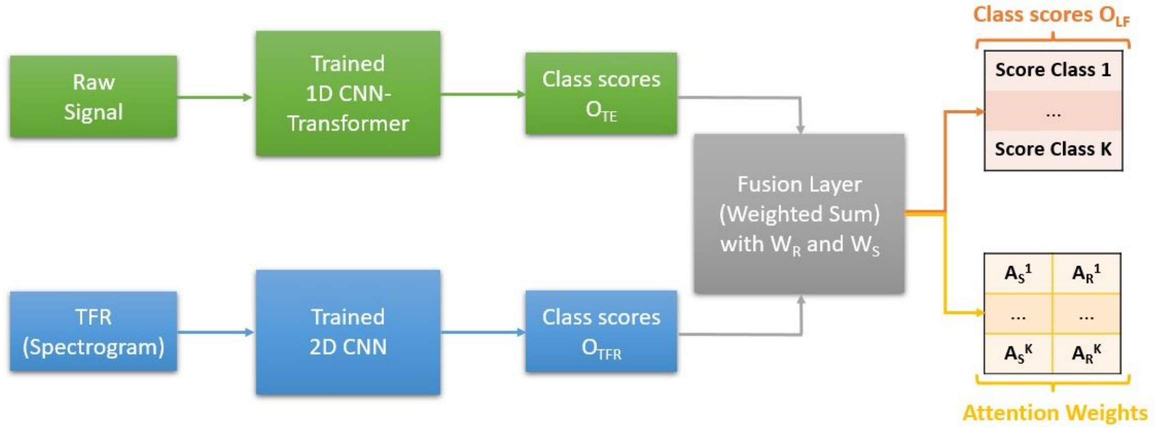

Figure 1: General pipeline of the proposed late fusion method. The green branch corresponds to the 1D-CNN-Transformer model extracting features from the raw signal. The blue branch corresponds to the 2D CNN model extracting features from the TFR. $\mathbf{W_R}$ and $\mathbf{W_S}$ are the raw signal and spectrogram attention weights for classification, respectively. The same subscript convention (S and R) is used for the normalized attention weights, $\mathbf{A_R}$ and $\mathbf{A_S}$

tained after the CNN blocks. The TE exploits the temporal information of the embeddings thanks to a sinusoidal positional encoding and learn hidden representations using an attention mechanism. The obtained representation, denoted $\mathbf{H_{TE}}$, can be combined with hidden features from other representations of the raw signal, or it can be fed to a specific classifier to perform the classification. If classification is performed, we denote by $\mathbf{O_{TE}} \in \mathbb{R}^{K\times 1}$ the classification scores, where K is the number of classes that we want to classify, and we feed the FC layers with a class token extracted from $\mathbf{H_{TE}}$ as in (Dosovitskiy et al., 2020).

3.2. 2D CNN model

Let us denote the magnitude spectrogram in logarithmic scale by $\mathbf{S} = [\mathbf{S_1}, ..., \mathbf{S_2}] \in \mathbb{R}^{F\times M}$, where F is the number of frequency bins and M is the number of time bins.

To extract features from the TFR, we use a conventional 2D CNN architecture (each spectrogram is processed as an image). A summary of the used architecture can be found in Figure 3. The model is composed of four convolutional blocks, each block composed of a 2D convolutional filter, a batch normalization layer, a leaky ReLU activation, and a pooling layer. The obtained representation, denoted as $\mathbf{H_{TFR}}$, can be combined with a hidden feature of the raw signal, or it can be fed through one FC layer to do classification. If classification is done, we denote by $\mathbf{O_{TFR}} \in \mathbb{R}^{K\times 1}$ the output classification scores.

3.3. Late fusion module

The first fusion method that we introduce is the late fusion method, which takes the output of two classification models and combines them using learnable and interpretable attention weights. Let us denote by $\mathbf{W_R} \in \mathbb{R}^{K\times 1}$ the attention weight vector associated with

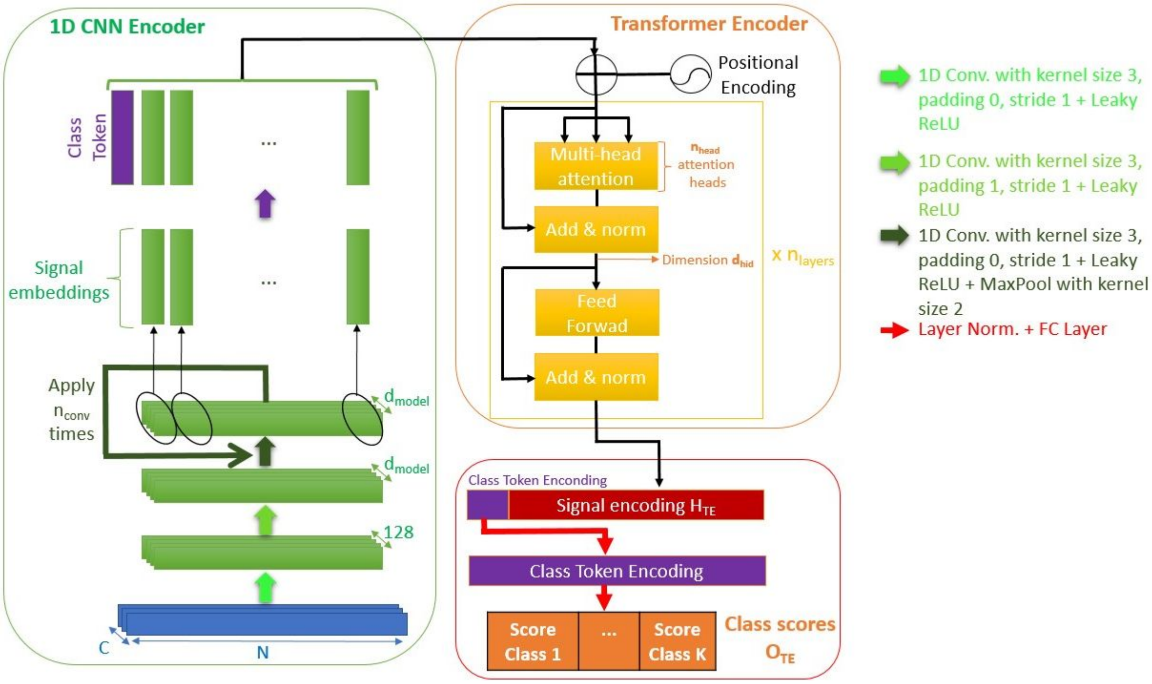

Figure 2: Hybrid 1D CNN-Transformer architecture

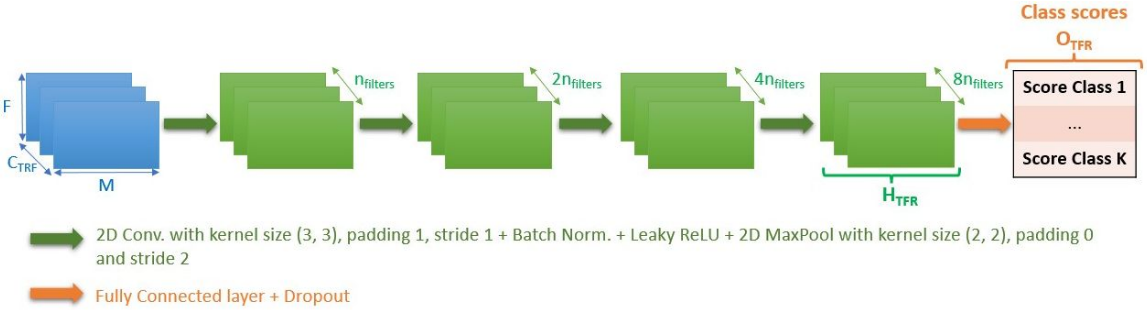
Class scores
O_{TFR}
F
C_{TRF}
M
$n_{filters}$
$2n_{filters}$
$4n_{filters}$
$8n_{filters}$
H_{TFR}
Score Class 1
...
Score Class K
2D Conv. with kernel size (3, 3), padding 1, stride 1 + Batch Norm. + Leaky ReLU + 2D MaxPool with kernel size (2, 2), padding 0 and stride 2
Fully Connected layer + Dropout

Figure 3: 2D CNN architecture used for classification using as input a time-frequency representation.

the raw signal representation $\mathbf{H_{TE}}$. Similarly, let's denote by $\mathbf{W_S} \in \mathbb{R}^{K\times 1}$ the attention weight vector associated with the spectrogram representation $\mathbf{H_{TFR}}$. We compute the final classification scores $\mathbf{O_{LF}}$ (late fusion) as follows:

$$\mathbf{O_{LF}} = \mathbf{W_R} \odot \mathbf{O_{TE}} + \mathbf{W_S} \odot \mathbf{O_{TFR}} \tag{1}$$

where $\odot$ represents the Hadamard product.

The weights $\mathbf{W_R}$ and $\mathbf{W_S}$ are learned using backpropagation. To obtain more interpretable weights, after the learning process is completed, we transform the weights into scores by applying a softmax function.

$$\mathbf{A_S} = softmax(\mathbf{W_S}) \tag{2}$$

$$\mathbf{A_R} = softmax(\mathbf{W_R}) \tag{3}$$

The element $\mathbf{A_R}^i$ represents the importance of the raw signal representation for the classification score of the class i of the classification model. Similarly, $\mathbf{A_S}^j$ represents the importance of the spectrogram representation for the classification score of class j.

3.4. Intermediate fusion modules

In addition to weighted late fusion, we tested three types of intermediate fusion: concatenation, sum and weighted attention sum.

First, since $\mathbf{H_{TE}}$ and $\mathbf{H_{TFR}}$ do not live in spaces of the same dimension, we project them into spaces of equal dimension (64) using a FC layer for each one. This gives us two new representations, $\tilde{\mathbf{H}}_{\mathbf{TE}}$ and $\tilde{\mathbf{H}}_{\mathbf{TFR}}$.

Then, we combine the obtained representation using one of the aforementioned methods. We denote $\mathbf{H_{cat}}$ the concatenated feature, $\mathbf{H_{sum}}$ the summed feature, and $\mathbf{H_{att\_sum}}$ the weighted sum feature. They are obtained as follows:

$$\mathbf{H_{cat}} = \tilde{\mathbf{H}}_{\mathbf{TFR}} \oplus \tilde{\mathbf{H}}_{\mathbf{TE}} \tag{4}$$

$$\mathbf{H_{sum}} = \tilde{\mathbf{H}}_{\mathbf{TFR}} + \tilde{\mathbf{H}}_{\mathbf{TE}} \tag{5}$$

$$\mathbf{H_{att\_sum}} = \alpha \times \tilde{\mathbf{H}}_{\mathbf{TFR}} + \beta \times \tilde{\mathbf{H}}_{\mathbf{TE}} \tag{6}$$

where α, $\beta \in \mathbb{R}$ are learnable attention weights that indicate the global importance of each representation for the final decision of the model.

Finally, the obtained representation is passed through an FC layer of shape $64 \times K$ to perform the classification.

4. Data

To train and evaluate our proposed method, we used three medical datasets: a private Transcranial Doppler (TCD) dataset, called the HITS dataset (Vindas et al., 2022), and two public electrocardiograms (ECG) datasets from Physionet (Goldberger et al., 2000), the PTB (Bousseljot et al., 1995) and MIT-BIH (Moody and Mark, 2001) datasets.

4.1. HITS dataset

4.1.1. Data acquisition

TCD recordings were performed on 39 subjects (15 men, 19 women, and 5 unknown; median age 63, range 21 to 85, computed with the available information) of 11 different centers using an Atys Medical device (TCD-X Holter or WAKIe) with a 1.5 MHz robotized probe, allowing recordings between 30 and 180 minutes. Patients came from different care units (neurovascular and cardiovascular), have different pathologies (stenosis, patent foramen ovale or none), and were injected or not with different contrast agents (Sonovue and iodine-containing contrast agent). Additionally, the acquisition conditions were heterogeneous as some recordings were acquired during surgical procedures (transcatheter aortic valve implantation and atrial fibrillation ablation) and some not. What is more, according to the recommendations to monitor the MCA and to do emboli detection, we have the following acquisition information:

- Pulse repetition frequency: 4.4-6.2 kHz;
- Transmitted ultrasound frequency: 1.5 MHz;
- Insonation depth: $45 - 55\ mm$;
- Sample volume: $8 - 10\ mm^3$.

The dataset is composed of 403 artifacts, 569 gaseous emboli, 569 solid emboli, and 4 unknown HITS. Appendix A describes the distribution of HITS per subject. Furthermore, to train and evaluate the different models, we split the dataset into two subsets, one for training and one for testing, according to the subjects. In this way, the HITS of a given subject are either in the training set or in the testing set, but they cannot be in both sets.

4.1.2. Data pre-processing

The spectrograms were computed from the TCD signals using $n_{fft} = 128$ (length of the windowed signal after padding with zeros), a $n_{overlap} = 8$ (size of the overlap), and a Blackman window[2]. Then HITS were detected (9 dB threshold), resulting in 1545 extracted HITS distributed in three classes (artifact, gaseous emboli, and solid emboli), each of duration 250 ms. Moreover, in addition to the spectrogram, to each HITS we also associate a raw time dependent signal. These signals were normalized using the mean and standard deviation of the training set. Finally, the spectrograms of all HITS were transformed into images used to train the different models.

4.2. PTB and MIT-BIH datasets

As the HITS dataset is a private dataset, we also performed experiments using two publicly available heartbeat categorization datasets: PTB (Bousseljot et al., 1995) and MIT-BIH (Moody and Mark, 2001) from PhysioNet. Both datasets are composed of ECG lead-II recordings resampled at a frequency of 125 Hz. The PTB dataset focuses on the identification of myocardial infarction (two imbalanced classes, normal and abnormal, 14 552 samples) and the MIT-BIH dataset focusing on Arrhythmia classification (five imbalanced classes, 103 436 samples). We used the standardized version of both datasets presented in (Kachuee et al., 2018)[3]. In these versions, the ECG signals were segmented into heartbeats, denoised, and normalized. We computed the spectrograms from these signals using $n_{fft} = 32$, $n_{overlap} = 4$ and a Blackman window[2]. Finally, the authors also proposed a

[2] The choice of these parameters was motivated by a trade-off between model performance, model complexity, and available training data. Indeed, lower values of n_{fft} reduce the performance of the model while reducing the number of parameters. Higher values of n_{fft} increase performance (up to some threshold value) while increasing the number of parameters.

[3] We use the public available versions found in https://www.kaggle.com/datasets/shayanfazeli/heartbeat

training, validation, and testing splitting which was also used in this paper. Tables 7 and 8 describe the number of samples per class for the PTB and MIT-BIH datasets, respectively. For more details, the reader can refer to Kachuee et al. (2018) and Appendix B.

5. Experiments

We conduct two main experiments to evaluate the different aspects of our method. The first experiment evaluates the advantage of using multiple features to enhance the performances of a classification model. The second experiment compares different intermediate and late feature fusion methods.

5.1. Experiment 1: Advantage of using multiple features

The objective of this experiment is to compare the performance of the proposed models with and without the use of different initial representations to show the advantage of multiple initial representations. For each dataset, we train three models, one 1D CNN Transformer with class token using only the raw signal, one 2D CNN using only the spectrogram and one late fusion model with learnable attention weights using both representations (Hybrid). For this last model, we proceed as follows. We start by learning independently the classification scores of each representation by a classification task. Then, we freeze the weights of the trained models, and we learn the attention weights.

For the 1D CNN-Transformer model we used $n_{head} = 8$, $n_{layers} = 8$, $d_{hid} = 64$, $d_{model} = 128$, $d_{proj} = 10$, $dropout = 0.1$ and $n_{conv} = 2$ for the HITS dataset and $n_{conv} = 4$ for the PTB and MIT-BIH datasets. For the 2D CNN, we used a dropout probability of 0.2 and an initial number of convolutional filters of 256 for the HITS dataset and 32 for the PTB and MIT-BIH datasets.

Table 1 presents the training parameters of the different models. All models were trained using cross-entropy (CE) loss, with class weights to handle the imbalanced classes. Class weights were calculated using Scikit Learn (Pedregosa et al., 2011) and their approach is inspired by (King and Zeng, 2001). The 2D CNN and late fusion models were trained using the Adamax optimizer and the 1D-CNN-Transformer model was trained using Noam optimization (Vaswani et al., 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and 4000 warm-up steps.

All experiments were repeated 10 times and mean performances were compared using the Matthews correlation coefficient (MCC), the F1 score and the accuracy measured on the test set.

The results are shown in Table 2. First, we can see that for the three tested datasets and for all the metrics, the best performing approach is the one using both representations with late fusion and learned attention weights per representation and per class, with an increase in up to 4.30% in MCC, 4.27% in F1 score and 2.84% in accuracy. Secondly, for the HITS and PTB dataset we obtain state-of-the-art performances, outperforming the models in (Vindas et al., 2022) for the HITS dataset and (Ahmad et al., 2021) for PTB with a difference of 1.6% in terms of F1 score. Furthermore, our proposed method outperforms the manual classification performed by clinicians. Thirdly, we can see that globally, using multiple representations allow reducing the variability of the mean performance of the model, reducing in the best case by 0.45%. Moreover, for the MIT-BIH dataset, we get close performance to the multimodal image fusion (MIF) approach (98.4% against 98.6%)

Table 1: Training parameters for the different models. The hybrid model corresponds to the late fusion proposed method.

Model	**Dataset**	**Learning rate**	**Weight Decay**	**Batch Size**	**Epochs**
1D-CNN-Transformer	HITS	10^{-1}	10^{-4}	16	100
	PTB				150
	MIT-BIH				
2D CNN	HITS	10^{-5}	10^{-5}	4	40
	PTB	10^{-3}		16	30
	MIT-BIH				
Hybrid	HITS	10^{-2}	10^{-8}	16	15
	PTB	10^{-3}	10^{-2}		10
	MIT-BIH	3×10^{-4}	10^{-2}		

of (Ahmad et al., 2021) but we are unable to reach the performance of MFF (it outperforms our method by 1.3%). However, in section 6 we further discuss the relevance of the accuracy metric when dealing with imbalanced classes. Finally, table 4 shows the final attention weights for each class and each representation. We can see that based on the dataset and the class, one representation is more important than the other. This will be analyzed in Section 6.

5.2. Experiment 2: Influence of the fusion layer

The objective of this experiment is to highlight the advantages of late fusion with learnable attention weights compared to other fusion methods. To do this, we train in an end-to-end manner three more models per dataset, where the fusion is done at an intermediate state using equations 4, 5 6. Once the fusion is done, we pass the obtained representation to a set of two fully connected layers.

For the three new models, we used $n_{head} = 8$, $n_{layers} = 8$, $d_{hid} = 64$, $d_{model} = 128$, $d_{proj} = 10$, $p_{dropout} = 0.1$ and $n_{conv} = 2$, and an initial number of convolutional filters of 64 for the HITS dataset and 32 for PTB and MIT-BIH. The training parameters were the same for the new models; we used a learning rate of 10^{-4}, a weight decay of 10^{-4}, a number of epochs of 50, and a batch size of 8 for all the HITS models and for the MIT-BIH with summed representation, and 16 for the rest of the models. To optimize the models, we used Noam optimization with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and 4000 warm up steps. Additionally, we applied early stopping by selecting the model at the epoch with the maximum validation accuracy. All experiments were repeated 10 times and mean performances were compared using the Matthews correlation coefficient (MCC), the F1 score and the accuracy measured on the test set.

Results are shown in table 3. First, we can see that, for the three datasets, the late fusion method with attention weight outperforms the other intermediate fusion approaches, by a margin larger than 2.74% in terms of MCC except for the HITS dataset where the intermediate sum model performs similarly to the late fusion method. Second, we can see

Table 2: Results of experiment 1. The hybrid model corresponds to the late fusion proposed method.

Dataset	**Model**	**MCC**	**F1 score**	**Accuracy**
HITS	2D CNN (Vindas et al., 2022)	83.53 ± 2.98	85.68 ± 2.31	89.48 ± 2.06
	1D-CNN-Transformer	80.29 ± 1.83	85.36 ± 1.09	87.37 ± 1.23
	2D CNN	85.03 ± 3.06	86.88 ± 2.38	90.55 ± 2.12
	Hybrid	**89.33 ± 2.77**	**91.15 ± 1.97**	**93.39 ± 1.74**
PTB	Manual classification (Makimoto et al., 2020)	-	70 ± 7	67 ± 7
	(Liu et al., 2018)	-	96.36	96
	(Sharma et al., 2015)	-	95.91	96
	(Chen et al., 2018)	-	95.46	96.18
	(Ahamed et al., 2020)	-	-	97.66
	MIF (Ahmad et al., 2021)	-	95.96	98.4
	MFF (Ahmad et al., 2021)	-	98	99.2
	1D-CNN-Transformer	97.92 ± 0.28	98.96 ± 0.14	99.16 ± 0.11
	2D CNN	93.42 ± 2.27	96.66 ± 1.20	97.32 ± 0.91
	Hybrid	**99.29 ± 0.21**	**99.65 ± 0.10**	**99.71 ± 0.08**
MIT-BIH	(Zhao et al., 2017)	-	-	98.25
	(Huang et al., 2019)	-	-	99
	(He et al., 2021)	-	-	98.3
	(Qiao et al., 2020)	-	-	99.3
	(Li et al., 2019)	-	97.70	99.5
	MIF (Ahmad et al., 2021)	-	92.50	98.6
	MFF (Ahmad et al., 2021)	-	**98**	**99.7**
	1D-CNN-Transformer	93.17 ± 0.70	89.44 ± 0.99	97.87 ± 0.24
	2D CNN	91.26 ± 0.76	86.40 ± 1.39	97.34 ± 0.26
	Hybrid	**94.63 ± 0.29**	91.28 ± 0.54	98.37 ± 0.09

that, globally, the late fusion method considerably reduces the variability of the performance of the model (this is particularly true for the PTB and MIT-BIH datasets, where the variability can be reduced by 0.97%). Thirdly, comparing with the results of experiment 1 (Table 2) we can notice that the three types of intermediate fusion does not improve the performance with respect to the use of a single representation. Indeed, besides the intermediate sum model on the HITS dataset, all the other models have similar or even worse performances than their single-spectrogram counterpart (with an MCC degradation up to 1.3% in the PTB dataset). Finally, we conclude that the performance of the three intermediate fusion methods is very close and that none of them competes with the late fusion approach.

Table 3: Results of experiment 2. The hybrid model corresponds to the late fusion proposed method.

Dataset	**Fusion Type**	**MCC**	**F1 score**	**Accuracy**
HITS	Concat.	84.96 ± 2.54	86.37 ± 2.11	90.62 ± 1.65
	Sum	89.04 ± 1.98	90.23 ± 1.71	93.16 ± 1.29
	Weighted Sum	86.31 ± 2.80	87.73 ± 2.32	91.31 ± 1.92
	Hybrid	$\mathbf{89.33 \pm 2.77}$	$\mathbf{91.15 \pm 1.97}$	$\mathbf{93.39 \pm 1.74}$
PTB	Concat.	92.91 ± 2.61	96.42 ± 1.33	97.11 ± 1.05
	Sum	92.12 ± 2.33	96.02 ± 1.19	96.78 ± 0.99
	Weighted Sum	92.74 ± 2.01	96.35 ± 1.00	97.06 ± 0.81
	Hybrid	$\mathbf{99.29 \pm 0.21}$	$\mathbf{99.65 \pm 0.10}$	$\mathbf{99.71 \pm 0.08}$
MIT-BIH	Concat.	91.51 ± 0.79	86.93 ± 1.10	97.42 ± 0.27
	Sum	91.89 ± 0.47	87.50 ± 0.87	97.55 ± 0.15
	Weighted Sum	91.56 ± 0.72	86.70 ± 1.13	97.44 ± 0.24
	Hybrid	$\mathbf{94.63 \pm 0.29}$	$\mathbf{91.28 \pm 0.54}$	$\mathbf{98.37 \pm 0.09}$

6. Discussion

Experiment 1: Advantage of using multiple features The results of experiment 1 confirm the genericity of our method, as well as the interest in using our proposed method to improve the classification performance of a model in three different medical datasets. Our proposed method takes advantage of the complementarity of both representations, the raw signal focusing on the temporal context and the amplitude information, whereas the spectrogram focuses on the spectral information. Moreover, the results show the genericity of our method. Indeed, it was tested on three different datasets corresponding to three different tasks and showed the same behavior and great performances on the three datasets. This is one of the main advantages of our method, as it proposes to exploit two of the classical representations used for signal classification, instead of having to choose between one of them. Furthermore, this experiment also highlights another advantage of our method, the stability of the final classification. Indeed, besides for the HITS dataset, for the PTB and MIT-BIH dataset the use of both representations allowed to reduce the variability in the test MCC, F1 score and accuracy scores. This is particularly interesting in the medical field, where we need stable models capable of giving similar results independently of the randomness of the training procedure.

Furthermore, our method was able to achieve state-of-the-art performances on the HITS and PTB datasets. However, to do a more fair comparison with the method proposed in (Ahmad et al., 2021), we should compare other metrics such as MCC because we are dealing with highly imbalanced datasets (especially the PTB and MIT-BIH datasets). Moreover, our proposed model for the MIT-BIH dataset is smaller than the best performing model of (Ahmad et al., 2021) by a factor of 8 (9 259 427 against 1 159 840 trainable parameters), and achieves similar performance than the MIF method, which has around 2.5 times more parameters than our proposed method. By the same token, we can see that the performance on the HITS dataset is smaller than the ones obtained on PTB or MIT-BIH. This can be explained by two main reasons: the size of the dataset, the available temporal context, and the complexity of the task. Indeed, the HITS dataset has around 500 samples per class, whereas the PTB dataset has at least 5000 samples per class, and the MIT-BIH has 800 samples per class (minority class). Moreover, the duration of the PTB and MIT-BIH samples is around 1.44 s whereas for the HITS dataset it is of around 0.250 s (less than one cardiac cycle), which is around 5 times smaller. Finally, the emboli classification is more complex as even for a human expert, identifying some solid emboli from gaseous emboli or artifacts can be difficult (as the unknown samples of the dataset show it).

Furthermore, our method has three major drawbacks. First, the model is longer to train. Indeed, instead of training a single model, we need to start by training two independent models and then train a final classifier using the attention weights. This drawback can partially be solved by training in parallel the two initial models (the fine-tuning of the attention weights is relatively fast). Secondly, the method is harder to optimize. Indeed, we have three models to train, and each model has different hyperparameters that have to be optimized. Third, the multiple features late fusion model is heavier in terms of memory than single feature models as we increase the number of parameters. Indeed, the final late fusion hybrid models has 26 073 416 (HITS), 1 156 732 (PTB), and 1 159 840 (MIT-BIH) learnable parameters. Moreover, they use 19.87 G mult-adds (HITS) and 0.119 G mult-adds (PTB and MIT-BIH), the model size is of 302 MB for the HITS model and 7 MB for the ECG models and the mean inference time is smaller than 1 s (using Intel (R) Xeon (R) CPU E5-2650L v3 @ 1.80GHz and no GPU). Some solutions such as quantization, pruning, and Huffman encoding can considerably reduce the size of the models.

Experiment 2: Influence of the fusion layer The results of experiment 2 raise an important point: fusion does not always increase the performances of the models, and using a wrong fusion strategy can even reduce their performances. Indeed, in the PTB and MIT-BIH datasets, intermediate fusion leads to similar or even worse results than spectrogram-only representations. On the contrary, our fusion approach always increases the classification performances, outperforming the three other fusion methods by an important margin (up to 4% in terms of MCC and F1 score and 3% in terms of accuracy). This confirms that our method is able to exploit better than the other tested methods the complementarity of both representations thanks to the learned attention weights. The only exception is on the HITS dataset, since the intermediate sum approach achieves similar results to our proposed approach. However, in that case, the model is not interpretable with respect to the importance of each representation for the final decision of the model. Moreover, our approach allows to considerably reduce the variability on the PTB and the MIT-BIH datasets. This is not noticeable in the HITS dataset, which can be explained by two reasons. First, for the HITS dataset, the best performing feature is the spectrogram (contrary to PTB and MIT-BIH), which has the greater variability. Second, as the attention weights of table 4 show it, the final decision of the hybrid model is more based on the spectrogram representation than the raw signal. Therefore, the final variability of the model is more influenced by the variability of the spectrogram-only model than the one of the raw signal only model.

Table 4: Attention weights median values and mean absolute deviations for the late fusion model on the three used datasets

Dataset	**Class**	**Spectrogram**	**Raw Signal**
HITS	Artifact	0.46 ± 0.29	0.54 ± 0.29
	Gaseous emboli	0.65 ± 0.17	0.35 ± 0.17
	Solid emboli	0.71 ± 0.15	0.29 ± 0.15
PTB	Normal	0.49 ± 0.12	0.51 ± 0.12
	Abnormal	0.18 ± 0.10	0.82 ± 0.10
MIT-BIH	N	0.48 ± 0.01	0.52 ± 0.01
	S	0.50 ± 0.01	0.50 ± 0.01
	V	0.50 ± 0.01	0.50 ± 0.01
	F	0.49 ± 0.02	0.51 ± 0.02
	Q	0.50 ± 0.003	0.50 ± 0.003

This last point illustrates the interest of the attention weights for interpretability purposes. Indeed, our method offers interpretable attention weights for each representation and for each class, as showed in table 4. This can give interesting insights for the use of different modalities, even for annotation purposes. When we study the annotation weights of the HITS dataset, we see that for the artifact class both representations are equally important. However, for the solid emboli and gaseous emboli classes, the spectrogram modality is more important than the raw signal modality. This is consistent with the manual annotation process. Indeed, when an annotator labels HITS data, they start by seeing the spectrogram. In many cases, the spectrogram is discriminating enough to classify the sample. However, in some cases, the expert can hesitate and use the raw signal to eliminate doubt. For the PTB dataset, we can see that the raw signal is more useful to identify abnormal heartbeats than the spectrogram. However, the results indicate that, in case of doubt, the spectrogram can be helpful.

Finally, our method has another important advantage over the other presented fusion approaches: it is easier to optimize. Indeed, we just need to optimize each single feature model independently, and then fine-tune the attention weights, which is not a difficult task. For the intermediate fusion methods, we add FC layers which add extra parameters and hyperparameters, making the model harder to optimize and heavier in terms of memory. Nevertheless, to limit the negative impact of poorly performing single feature models we plan to further improve our method with an end-to-end training strategy, for instance via iterated losses (Tjandra et al., 2020) or direct end-to-end training.

7. Conclusion

In this paper, we proposed a novel CNN-Transformer model based on multi-feature extraction and learnable representation attention weights per class to perform classification with raw signals and TFRs. Instead of choosing one fixed initial representation of the signal, our method proposed to exploit two complementary representations: the raw signal (temporal information) and the spectrogram (spectral information). We pass these two representations to two different models, a 1D CNN Transformer for the raw signal and a 2D CNN for the spectrogram. Then, we fuse the output of each model using a late fusion mechanism with learnable and interpretable weights. These weights attribute an importance of each representation for the final classification score of each class. Extensive experiments in three different datasets demonstrate the effectiveness of our method, improving the classification performances up to 3% in terms of classification accuracy and up to 4% in terms of MCC and F1 score.

Acknowledgments

This work was carried out in the context of the CAREMB project funded by the Auvergne-Rhône-Alpes region, within the *Pack Ambition Recherche* program. This work was performed within the framework of the LABEX CELYA (ANR-10-LABX-0060) and PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

Suneil K. Aggarwal, Nicola Delahunty RN, Leon J. Menezes, Richard Perry, Bethany Wong, Markus Reinthaler, Muhiddin Ozkor, and Michael J. Mullen. Patterns of solid particle embolization during transcatheter aortic valve implantation and correlation with aortic valve calcification. *Journal of Interventional Cardiology*, 31(5):648–654, 2018. ISSN 08964327. doi: 10.1111/joic.12526. URL http://doi.wiley.com/10.1111/joic.12526. Number: 5.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1):4–31, may 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0966-6. URL https://doi.org/10.1007/s11263-016-0966-6.

Md. Atik Ahamed, Kazi Amit Hasan, Khan Fashee Monowar, Nowfel Mashnoor, and Md. Ali Hossain. Ecg heartbeat classification using ensemble of efficient machine learning approaches on imbalanced datasets. In *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, pages 140–145, 2020. doi: 10.1109/ICAICT51780.2020.9333534.

Zeeshan Ahmad, Anika Tabassum, Ling Guan, and Naimul Mefraz Khan. Ecg heartbeat classification using multimodal fusion. *IEEE Access*, 2021.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 05 2017. doi: 10.1109/TPAMI.2018.2798607.

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. 40(s1):317–318, 1995. doi: doi:10.1515/bmte.1995.40.s1.317. URL https://doi.org/10.1515/bmte.1995.40.s1.317.

Ginevra Castellano, Loic Kessous, and George Caridakis. *Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech*, pages 92–103. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-85099-1. doi: 10.1007/978-3-540-85099-1_8. URL https://doi.org/10.1007/978-3-540-85099-1_8.

Xianjie Chen, Zhaoyun Cheng, Sheng Wang, Guoqing Lu, Gaojun Xv, Qianjin Liu, and Xiliang Zhu. Atrial fibrillation detection based on multi-feature extraction and convolutional neural network for processing ecg signals. *Computer Methods and Programs in Biomedicine*, 202:106009, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2021.106009. URL https://www.sciencedirect.com/science/article/pii/S0169260721000845.

Yufei Chen, Huihui Chen, Ziyang He, Cong Yang, and Yangjie Cao. Multichannel lightweight convolution neural network for anterior myocardial infarction detection. In *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 572–578, 2018. doi: 10.1109/SmartWorld.2018.00119.

Georges A. Darbellay, Rebecca Duff, Jean-Marc Vesin, Paul-André Despland, Dirk W. Droste, Carlos Molina, Joachim Serena, Roman Sztajzel, Patrick Ruchat, Theodoros Karapanayiotides, Afksendyios Kalangos, Julien Bogousslavsky, Erich B. Ringelstein, and Gérald Devuyst. Solid or gaseous circulating brain emboli: Are they separable by transcranial ultrasound? *Journal of Cerebral Blood Flow & Metabolism*, 24(8):860–868, 2004. doi: 10.1097/01.WCB.0000126235.54306.FA. URL https://doi.org/10.1097/01.WCB.0000126235.54306.FA. PMID: 15362716.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Israel D. Gebru, Silèye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1086–1099, 2018. doi: 10.1109/TPAMI.2017.2648793.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Blaise Kevin Guepie, Matthieu Martin, Victor Lacrosaz, Marilys Almar, Benoit Guibert, and Philippe Delachartre. Sequential emboli detection from ultrasound outpatient data. *IEEE Journal of Biomedical and Health Informatics*, 23(1):334–341, 2019. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2018.2808413. URL https://ieeexplore.ieee.org/document/8300318/. Number: 1.

Runnan He, Yang Liu, Kuanquan Wang, Na Zhao, Yongfeng Yuan, Qince Li, and Henggui Zhang. Automatic detection of qrs complexes using dual channels based on u-net and bidirectional long short-term memory. *IEEE Journal of Biomedical and Health Informatics*, 25:1052–1061, 2021.

Chiori Hori, Takaaki Hori, Gordon Wichern, Jue Wang, Teng-Yok Lee, Anoop Cherian, and Tim K. Marks. Multimodal attention for fusion of audio and spatiotemporal features for video description. In *CVPR Workshops*, 2018.

Jingshan Huang, Binqiang Chen, Bin Yao, and Wangpeng He. Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network. *IEEE Access*, 7:92871–92880, 2019. doi: 10.1109/ACCESS.2019.2928017.

Jikun Jin, Sihao Yang, Bingmei Zhao, Lizhu Luo, and Wai Lok Woo. Attention-block deep learning based features fusion in wearable social sensor for mental wellbeing evaluations. *IEEE Access*, 8:1–1, 05 2020. doi: 10.1109/ACCESS.2020.2994124.

Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. In *2018 IEEE international conference on healthcare informatics (ICHI)*, pages 443–444. IEEE, 2018.

Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1005. URL https://aclanthology.org/D14-1005.

Jin-Gyeom Kim and Bowon Lee. Appliance classification by power signal analysis based on multi-feature combination multi-layer lstm. *Energies*, 12(14), 2019. ISSN 1996-1073. doi: 10.3390/en12142804. URL https://www.mdpi.com/1996-1073/12/14/2804.

Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9: 137–163, Spring 2001.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Honglak Lee, Peter Pham, Yan Largman, and Andrew Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/a113c1ecd3cace2237256f4c712f61b5-Paper.pdf.

Feiteng Li, Jiaquan Wu, Menghan Jia, Zhijian Chen, and Yu Pu. Automated heartbeat classification exploiting convolutional neural network with channel-wise attention. *IEEE Access*, 7:122955–122963, 2019. doi: 10.1109/ACCESS.2019.2938617.

Wenhan Liu, Mengxin Zhang, Yidan Zhang, Yuan Liao, Qijun Huang, Sheng Chang, Hao Wang, and Jin He. Real-time multilead convolutional neural network for myocardial infarction detection. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1434–1444, 2018. doi: 10.1109/JBHI.2017.2771768.

Hisaki Makimoto, Moritz Höckmann, Tina Lin, David Glöckner, Shqipe Gerguri, Lukas Clasen, Jan Schmidt, Athena Assadi-Schmidt, Alexandru Bejinariu, Patrick Müller, Stephan Angendohr, Mehran Babady, Christoph Brinkmeyer, Asuka Makimoto, and Malte Kelm. Performance of a convolutional neural network derived from an ecg database in recognizing myocardial infarction. *Scientific Reports*, 10:8445, 05 2020. doi: 10.1038/s41598-020-65105-x.

Hugh S. Markus and Martin Punter. Can transcranial doppler discriminate between solid and gaseous microemboli?: Assessment of a dual-frequency transducer system. *Stroke*, 36(8):1731–1734, 2005. ISSN 0039-2499, 1524-4628. doi: 10.1161/01.STR.0000173399.20127.b3. URL https://www.ahajournals.org/doi/10.1161/01.STR.0000173399.20127.b3. Number: 8.

Gary Mckeown, Michel Valstar, Roddy Cowie, and Maja Pantic. The semaine corpus of emotionally coloured character interactions. pages 1079–1084, 07 2010. doi: 10.1109/ICME.2010.5583006.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2772–2778. AAAI Press, 2016.

G.B. Moody and R.G. Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. doi: 10.1109/51.932724.

Meinard Müller. Dynamic time warping. *Information Retrieval for Music and Motion*, 2:69–84, 01 2007. doi: 10.1007/978-3-540-74048-3_4.

Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Vij, and Jonathan Rubin. A wide and deep transformer neural network for 12-lead ecg classification. In *2020 Computing in Cardiology*, pages 1–4, 2020. doi: 10.22489/CinC.2020.107.

Dung Nguyen, Duc Thanh Nguyen, Rui Zeng, Thanh Thi Nguyen, Son Tran, Thin Khac Nguyen, S. Sridharan, and Clinton Fookes. Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2021. doi: 10.1109/TMM.2021.3063612.

Consensus Committee of the Ninth International Cerebral Hemodynamic Symposium. Basic identification criteria of doppler microembolic signals. *Stroke*, 26(6):1123, 1995.

Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2344, 2014. doi: 10.1109/CVPR.2014.299.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1303. URL https://aclanthology.org/D15-1303.

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019. doi: 10.1109/JSTSP.2019.2908700.

Fengjuan Qiao, Bin Li, Youmei Zhang, Hongli Guo, Wei Li, and Shuwang Zhou. A fast and accurate recognition of ecg signals based on elm-lrf and blstm algorithm. *IEEE Access*, 8:71189–71198, 2020. doi: 10.1109/ACCESS.2020.2987930.

Waseem Rawat and Zenghui Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 29(9):2352–2449, 09 2017. ISSN 0899-7667. doi: 10.1162/neco_a_00990. URL https://doi.org/10.1162/neco_a_00990.

Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. In Juergen Gall, Peter Gehler, and Bastian Leibe, editors, *Pattern Recognition*, pages 209–221, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24947-6.

L. N. Sharma, R. K. Tripathy, and S. Dandapat. Multiscale energy and eigenspace approach to detection and localization of myocardial infarction. *IEEE Transactions on Biomedical Engineering*, 62(7):1827–1837, 2015. doi: 10.1109/TBME.2015.2405134.

Abdelghani Tafsast, Karim Ferroudji, Mohamed Laid Hadjili, Ayache Bouakaz, and Nabil Benoudjit. Automatic microemboli characterization using convolutional neural networks and radio frequency signals. In *2018 International Conference on Communications and Electrical Engineering (ICCEE)*, pages 1–4. IEEE, 2018. ISBN 978-1-72810-112-5. doi: 10.1109/CCEE.2018.8634521. URL https://ieeexplore.ieee.org/document/8634521/.

Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '12, page 275–284, Goslar, DEU, 2012. Eurographics Association. ISBN 9783905674378.

Leslie Tiong, Seong Tae Kim, and Yong Ro. Implementation of multimodal biometric recognition via multi-feature deep learning networks and feature fusion. *Multimedia Tools and Applications*, 78, 08 2019. doi: 10.1007/s11042-019-7618-0.

Andros Tjandra, Chunxi Liu, Frank Zhang, Xiaohui Zhang, Yongqiang Wang, Gabriel Synnaeve, Satoshi Nakamura, and Geoffrey Zweig. Deja-vu: Double feature presentation and iterated loss in deep transformer networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6899–6903, 2020. doi: 10.1109/ICASSP40776.2020.9052964.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Yamil Vindas, Blaise Kévin Guépié, Marilys Almar, Emmanuel Roux, and Philippe Delachartre. Semi-automatic data annotation based on feature-space projection and local quality metrics: an application to cerebral emboli characterization. *Medical Image Analysis*, page 102437, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102437. URL https://www.sciencedirect.com/science/article/pii/S1361841522000883.

Sean Wallace, Gaute Døhlen, Henrik Holmstrøm, Christian Lund, and David Russell. Cerebral microemboli detection and differentiation during transcatheter closure of atrial septal defect in a paediatric population. *Cardiology in the Young*, 25(2):237–244, 2015. ISSN 1047-9511, 1467-1107. doi: 10.1017/S1047951113002072. URL https://www.cambridge.org/core/product/identifier/S1047951113002072/type/journal_article. Number: 2.

Ting Yao, Farong Gao, Qizhong Zhang, and Yuliang Ma. Multi-feature gait recognition with dnn based on semg signals. *Mathematical Biosciences and Engineering*, 18:3521–3542, 05 2021. doi: 10.3934/mbe.2021177.

Yong Zhao, Xueting Yin, and Yannan Xu. Electrocardiograph (ecg) recognition based on graphical fusion with geometric algebra. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 1482–1486, 2017. doi: 10.1109/ICISCE.2017.309.

Yinghui Zhu and Yuzhen Jiang. Optimization of face recognition algorithm based on deep learning multi feature fusion driven by big data. *Image and Vision Computing*, 104:104023, 2020. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2020.104023. URL https://www.sciencedirect.com/science/article/pii/S0262885620301554.

Appendix A. Distribution of HITS per class and per subject.

Table 5: Distribution of the HITS per class and per subject (subjects 0 to 19). The HITS are classified using three classes: artifacts, solid emboli and gaseous emboli. Some HITS are classified as unknown, but they are not used to train or evaluate the classification models. Indeed, in some cases, an expert is not able to annotate a HITS. This happens particularly when a HITS can be a solid or gaseous emboli, or when there is doubt between a small intensity solid emboli and an artifact.

Subject ID	Artifacts	Solid emboli	Gaseous embolus	Unknown	Total
0	15	0	123	1	139
1	1	24	3	0	28
2	0	0	72	0	72
3	46	11	0	0	57
4	0	1	0	0	1
5	0	2	0	0	2
6	48	0	0	0	48
7	0	3	0	0	3
8	0	56	0	0	56
9	54	1	0	0	55
10	0	0	4	0	4
11	0	1	0	0	1
12	0	0	15	0	15
13	0	0	76	0	76
14	0	2	0	0	2
15	46	5	0	0	51
16	0	3	0	0	3
17	4	14	0	0	18
18	0	2	0	0	2
19	0	0	54	0	54

Table 6: Distribution of the HITS per class and per subject (subjects 20 to 38). The HITS are classified using three classes: artifacts, solid emboli and gaseous emboli. Some HITS are classified as unknown, but they are not used to train or evaluate the classification models. Indeed, in some cases, an expert is not able to annotate a HITS. This happens particularly when a HITS can be a solid or gaseous emboli, or when there is doubt between a small intensity solid emboli and an artifact.

Subject ID	Artifacts	Solid emboli	Gaseous embolus	Unknown	Total
20	0	0	7	0	7
21	0	20	0	0	20
22	1	0	0	0	1
23	0	17	0	0	17
24	0	1	0	0	1
25	0	1	0	0	1
26	0	1	0	0	1
27	0	45	6	0	51
28	48	268	2	0	318
29	0	42	181	3	226
30	0	0	7	0	7
31	0	24	0	0	24
32	4	7	1	0	12
33	48	0	0	0	48
34	34	0	0	0	34
35	0	17	0	0	17
36	15	1	0	0	16
37	0	0	4	0	4
38	39	0	14	0	53

Appendix B. Number of samples per class for the PTB and MIT-BIH datasets.

Table 7: Number of samples per class in the PTB dataset.

Class	**Number of samples**
Normal	10506
Abnormal	4046

Table 8: Number of samples per class in the MIT-BIH dataset. Each of the classes regroups a set of abnormal heartbeats. To have the exact correspondence, see (Ahmad et al., 2021).

Class	**Number of samples**
N	90 589
S	2 779
V	7 226
F	803
Q	8 039

Appendix C. Interpretability of the model prediction using Integrated Gradients

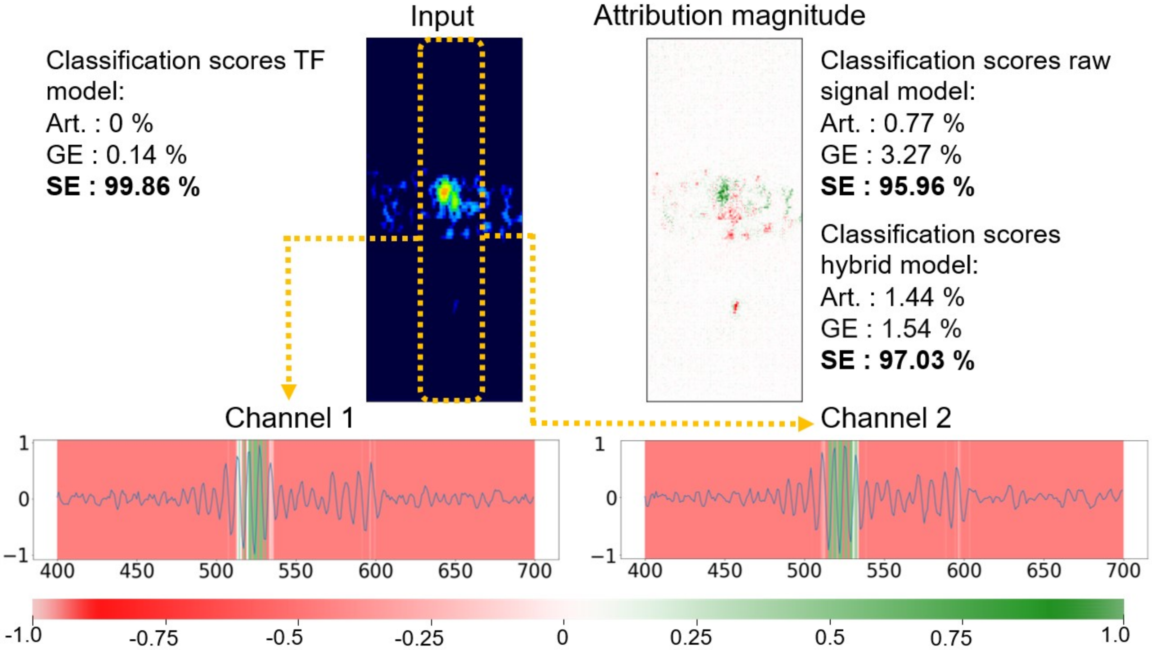

Figure 4: Integrated gradients' attribution maps of the multi-feature proposed model for a well classified HITS with respect to the true class (solid embolus). The attribution values were computed using the Python library Captum (Kokhlikyan et al., 2020). The single feature and multi-feature models agree on the class of the HITS. Globally, the model focuses on the maxima of the inputs. The model focuses on the high intensity zones in the spectrogram, with a special focus on the HITS. Some high intensity zones corresponding to the blood flow around the HITS disrupt the model for the prediction of the solid embolus class. This can be explained by the fact that gaseous embolus tend to have an elongated shape, so this high intensity zones around the HITS can confuse the model in favor of the gaseous embolus class. If we focus on the signal, we can see that the model focuses also on the HITS for both channels, but some attention is also given to the event just after, disrupting it towards the gaseous embolus class.

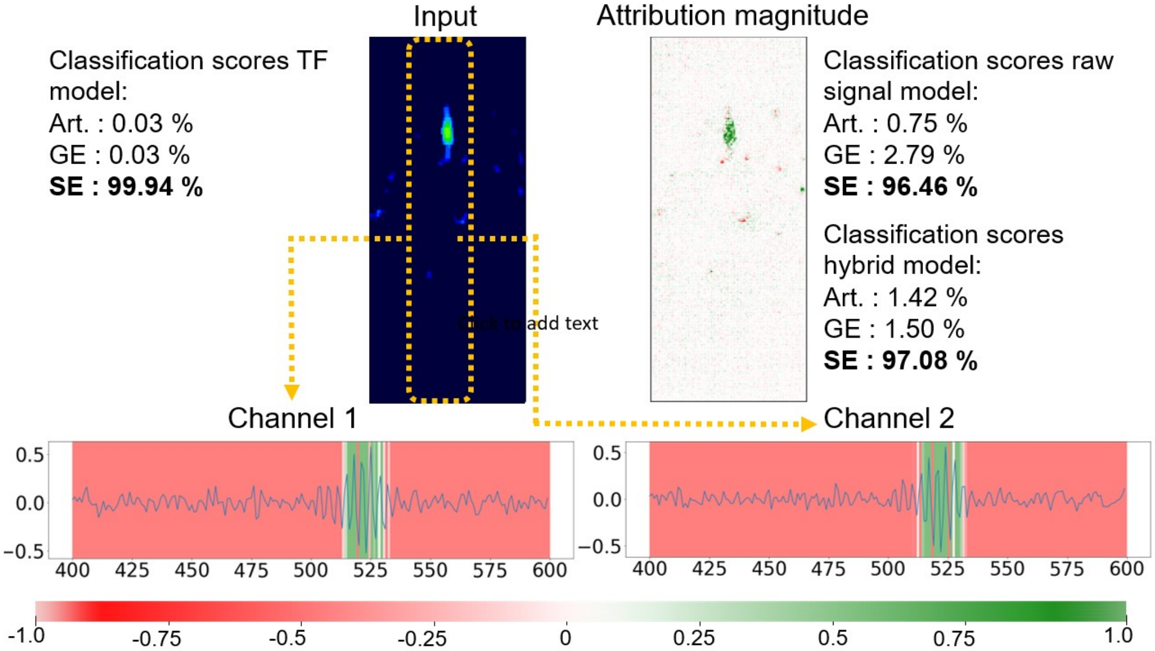

Figure 5: Integrated gradients' attribution maps of the multi-feature proposed model for a well classified HITS with respect to the true class (solid embolus). The attribution values were computed using the Python library Captum (Kokhlikyan et al., 2020). The single feature and multi-feature models agree on the class of the HITS. Globally, the model focuses on the maxima of the inputs. The model focuses on the high intensity zones in the spectrogram, with a special focus on the HITS. Contrary to the example in figure 4, the model focuses mainly in the HITS, with no blood flow disrupting the prediction. That is why we see a slight increase in the solid classification outputs of the raw signal model and hybrid model.

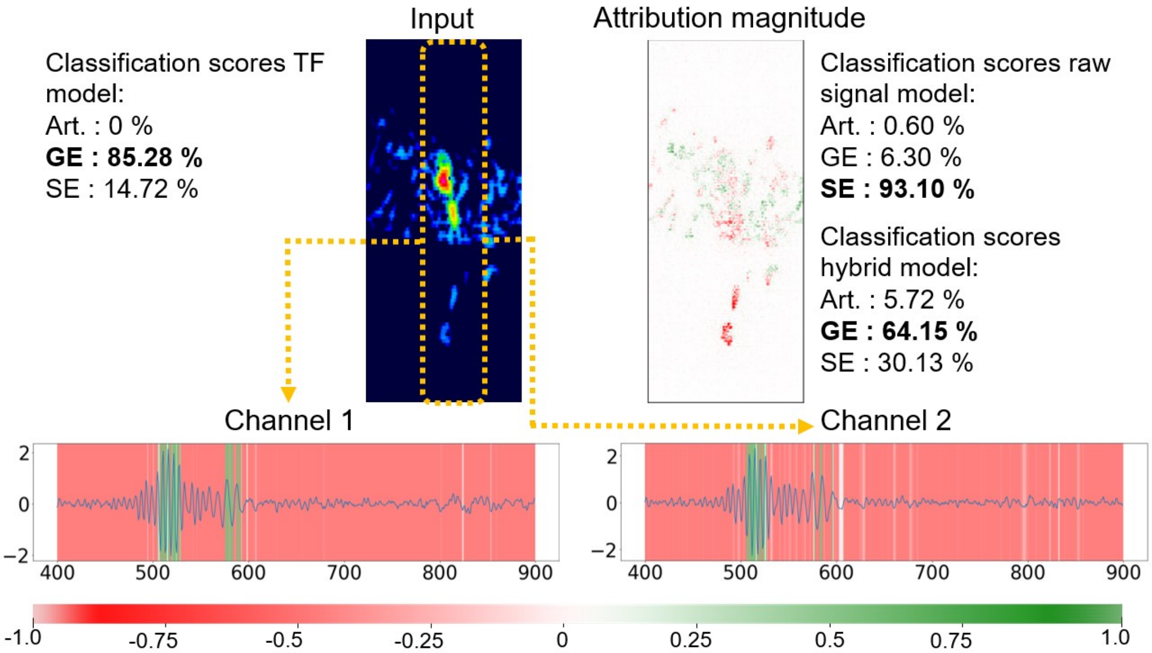

Figure 6: Integrated gradients' attribution maps of the multi-feature proposed model for a misclassified HITS with respect to the true class (solid embolus). attribution values were computed using the Python library Captum (Kokhlikyan et al., 2020). The single feature and multi-feature models do not agree on the class of the HITS. The model using only the spectrogram, misclassifies the HITS as a gaseous embolus, whereas the raw signal model accurately identifies the true class of the HITS. However, due to the attention weights, the classification of the final hybrid model is erroneous as the TF representation is more important than the raw signal representation. Nevertheless, we see that the final classification outputs of the hybrid model are not as confident as the outputs of the single feature models. Moreover, if we analyze the attribution maps, we can understand the decision of the different models. On the one hand, from the spectrogram attribution map, we can see that, for the prediction of the solid embolus class, the model does not use the HITS itself (which corresponds to the high red intensity zone in the middle of the spectrogram). What is more, the high yellow intensity zone under the HITS, as well as the high red intensity zones at the bottom of the spectrogram, disrupts its prediction. The former disrupts its prediction towards the gaseous embolus class as gaseous embolus often have an elongated shape. The latter disrupts the prediction towards the artifact class as many artifacts are symmetric, and this zone corresponds to the symmetric part of the HITS. From the other hand, the raw signal focuses well on the HITS itself, even though the event after it (high yellow intensity zone under the HITS in the spectrogram) is also used by the model, specially in the first channel. This is what mainly disrupts the classification of the signal model towards the gaseous embolus class.