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 transform it with 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 treat the raw signal and a 2D CNN for the TFR. Finally, we train a classifier combining 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 performances with respect to the use of a single feature or other fusion methods. Moreover, it achieves state-of-the-art results on the HITS and PTB datasets with a classification accuracy of $93,4\%$ and $99,7\%$ respectively, and it achieves great performances on the MIT-BIH dataset with an accuracy of $98,4\%$. What is more, our fusion method provides interpretable attention weights per class indicating the importance of each representation in the final decision of the classifier.

1. Introduction

Signals can be defined as encoded representations of physical phenomena. Images can be considered as signals as well as audios are. 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) (Krizhevsky et al., 2012; 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 (EGG) produce signals with a rich temporal dimension. As detailed hereafter, these signals can be used to detect pathologies such as patent foramen ovale (TCD) and arrhythmia (ECG) or they can be used for prevention purposes such as stroke prevention (TCD).

From a clinical point of view, stroke is one of the leading causes of death and disability in the world (Donkor, 2018). It can be caused by the blockage of cerebral arteries by cerebral emboli (Wallace et al., 2015) which are gaseous or solid particles that can circulate in the cerebral bloodstream. Several techniques such as computed tomography (CT), magnetic resonance imaging (MRI), and TCD ultrasound can be used to detect emboli (Wallace et al., 2015) in order to prevent stroke. In this paper we focus on TCD monitoring as it is a non-expensive and relatively cheap technique to detect emboli via High Intensity Transcient Signals (HITS). Contrary to standard MRI or CT sequences, TCD generate time dependent signals as it follows the bloodflow in the middle cerebral artery (MCA) over relatively long periods of time.

Classical signal processing techniques extract spectral and handcrafted features from the signals to do classification (Purwins et al., 2000, 2019). More recent approaches use deep learning techniques to automatically extract features from the signals or their time-frequency representations (TFRs). To exploit the temporal context of time-dependent signals, different models can be used such as 1D CNNs (Nguyen et al., 2021; Dieleman and Schrauwen, 2014), Recurrent Neural Networks (RNNs) (Nguyen et al., 2021; Hori et al., 2018) or Convolutional Deep Belief Networks (Lee et al., 2009).

One of the main difficulties when manipulating time-dependent signals with deep learning models is the choice of the optimal representation to use for the task to solve. Often, TFRs are used instead of the raw signal (Chaurasiya, 2020; Park and Yoo, 2020; Gong et al., 2021; Natarajan et al., 2020), even though the raw signal can give valuable and complementary information of the studied phenomenon (of the Ninth International Cerebral Hemodynamic Symposium, 1995). Moreover, some works propose to combine different features and/or representations (Kim and Lee, 2019; Yao et al., 2021; Chen et al., 2021; Jin et al., 2020) and the optimal way of combining them can have an important impact of the performance of the models (Ahmad et al., 2021). In these works, fusion is done by concatenation or majority vote (at different levels) using pre-computed representations from the raw signal. None of these works directly use the raw signal and only Jin et al. (2020) use attention weights to keep the final classification intepretable. However, it is not always clear how much a certain representation is useful for the decision of the final model, which make the model less interpretable.

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 weights. First, we compute the magnitude spectrogram of the raw signal in a logarithmic scale. Then, we feed the raw signal to an hybrid 1D CNN Transformer model and we feed the spectrogram to a 2D CNN model. This allows to extract hidden features from two different representations of the signal, one which focuses on the temporal characteristics (raw signal) and one which focuses on the spectral characteristics (spectrogram). The 1D CNN Transformer starts by extracting hidden features from the raw signal using a 1D CNN, and then a Transformer encoder exploits the temporal context thanks to its positional encoding mechanism and its attention mechanism. Afterwards, the output of the hybrid 1D CNN Transformer and the 2D CNN model are combined using learnable attention weights per modality and per class. This allow to obtain scores associated to the importance of each modality to the prediction probability of each class of the final model.

Our main contribution can be summarized as follows :

- A novel a 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 on 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 exploit directly the raw temporal dependent signal and/or both representations. In this work we focused on the use of both types of representations (temporal and spectral), as we found that it benefits the model performances 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 a nutshell, our method pushes further the deep model capabilities to exploit time dependent medical signals while maintaining the predictions interpretable.

2. Related Work

2.1. Multimodal learning

Multimodal learning has been an important topic of research in the past years (Baltrusaitis et al., 2017). The idea behind multimodal learning is to exploit the complementary information of different representations of a phenomenon in order to solve a task. Indeed, in many cases different modalities give different points of view which are complementary and can help to improve the performances of a model (Akbari et al., 2021). Baltrusaiti set al. (Baltrusaitis et al., 2017) establish a taxonomy of the different challenges in multimodal learning. We are going to focus on two challenges: multimodal representation and multimodal fusion.

Different representations can be created from multiple modalities: joint representations and coordinated representations. From 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; Wang et al., 2016). Other works use unsupervised learning techniques such as autoencoders (AEs) to extract features from each modality and the fuse the obtained representations with another AE model (Müller, 2007). Other models such as deep belief networks and deep Boltzmann machines have been used to extract a common representation from different modalities (Khapra et al., 2010; Socher et al., 2013, 2014). On the other hand, instead of creating a common representation of the different modalities, one can create individual representations coordinated between them in order to satisfy some constraint. Some models coordinate different representations by minimizing the distance between the representations of different modalities (Kiela and Bottou, 2014; Wöllmer et al., 2013). Other models try to enforce some structure on the representations of the different modalities thought different constraints such as order constraints (Taylor et al., 2012) and correlation (Poria et al., 2015; Shariat and Pavlovic, 2011).

Multimodal fusion is strongly linked to multimodal representation but it is not limited to the combination of the representations of different modalities (Baltrusaitis et al., 2017). Three fusion techniques commonly used are: early, intermediate and late fusion. Early fusion consists on combining the different modalities before the model. Castellano et al.

(Castellano et al., 2008) combine features extracted from face, body and speech data before feeding them through a Bayesian classifier and improving the classification performances of the model with respect to a single modality model. The main advantage of early fusion is that it is easy to implement and it allows to exploit the correlation between low level features of the available modalities (Baltrusaitis et al., 2017)]. Intermediate fusion combines the representations of different modalities after feeding them to the model but before taking the decision. Hori et al. (Hori et al., 2018) used an encoder-decoder RNN architecture and an attention fusion mechanism in the intermediate layers of the model to combine audio and video features for video description generation. Ortega et al. (Ortega et al., 2019) extracted features from audio, video and text data using DNN, then fuse the intermediate representation by concatenation, and finally do emotion recognition by feeding this joint representation to another DNN. Akbari et al. (Akbari et al., 2021) used a multimodal self-supervised Tansformer to exploit video, audio and text information on different tasks such as video action recognition, audio event classification and zero-shot retrieval. They proposed to extract hidden representation from each modality using a Transformer encoder, and then create different common representation spaces with different granularities using a contrastive loss. Late fusion takes different models trained with different modalities and combines their outputs. Different approaches allows to combine the outputs of different models such as 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 mainly focuses 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, in the past years different works have focused on ways of combining different representations coming from a single modality in order to improve the performances of a model for a given task.

In computer vision, several papers have tried to combine different features and/or representations of a single modality in order to enhance the performances of different models (Zhu and Jiang, 2020; Mao et al., 2020; Tiong et al., 2019; Wang et al., 2017). Zhang et Jiang (Zhu and Jiang, 2020) combined global and local features from face images in order to do face recognition. They used two-dimensional principal component analysis (2DPCA) to extract global features from the images and local binary patterns (LBP) to extract local features which were then fused and passed through a CNN. Similarly, Mao et al. (Mao et al., 2020) extracted color components using iterative RELIEF (a feature selction method) which were then passed through a CNN to extract hidden features. The obtained features were then combined by concatenation, compressed using PCA and passed through a support vector machine classifier. Tiong et al. (Tiong et al., 2019) extract different features from images such as histogram of gradients, LBP, and entropy texture to do face recognition. They used different CNN models to extract hidden representations from the different features and fuse the obtained representation in an intermediate layer using concatenation, averaging and max selection. They then passed 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 (Ahmad et al., 2021; Yao et al., 2021; Chen et al., 2021; Ertugrul et al., 2021; Liu, 2021; Feng et al., 2020; Jin et al., 2020; Kim and Lee, 2019). Kim et Lee (Kim and Lee, 2019) used a concatenation of three TFRs (spectrogram, mel-spectrogram and MFCC) with an LSTM to classify power signals. Jin et al. (Jin et al., 2020) did emotion recognition by using a LSTM model to extract features from different MFCCs and a DNN to extract features from behavioural data. They mixed the Mel-frequency cepstral coefficients (MFCCs) features using a weighted concatenation, and then they mixed the obtained representations with the DNN behavioural features using another weighted concatenation. Liu (Liu, 2021) did specific emitter identification by extracting amplitude, phase and spectrum asymmetry characteristics from raw signals, concatenate them and passing the fused feature through a 1D CNN model. Chen et al. (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 of one heartbeat and the coherence spectrum characteristics of two adjacent heartbeats), passed them through a 1D CNN and combine the outputs of the models by majority voting. Yao et al. (Yao et al., 2021) extract TFR from sEMG signals and mixed them by concatenation (early fusion) before feeding them through a DNN to classify them. Ahmad et al. (Ahmad et al., 2021) did heartbeat categorization using ECG signals. They started by extracting three images from the raw signal: gramian angular field (GAF), recurrent plot (RP), and Markov transition field (MTF). They then proposed two approaches. The first one fuse these features before passing them through an AlexNet model (this approach is called multimodal image fusion or MIF). The second one pass each feature to an AlexNet model, and then fuse the extracted hidden features before given them to a SVM classifier (this approach is called multimodal feature fusion or MFF). By doing this, they achieve state-of-the-art performances on two heartbeat categorization datasets, PTB and MIT-BIH.

Finally, other fields such as information retrieval (Abdi et al., 2019, 2021) and bioinformatics (Wekesa et al., 2020) have used similar multi-feature fusion techniques to exploit complementary information of different representations of a single modality.

3. Methods

In this paper we propose a novel model for classification 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 specific feature extraction 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 from Natarajan et al. (Natarajan et al., 2020) and it is resumed in figure 2. The first blocks corresponds to 1D CNN blocks allowing to efficiently extract features from the raw signal thanks to

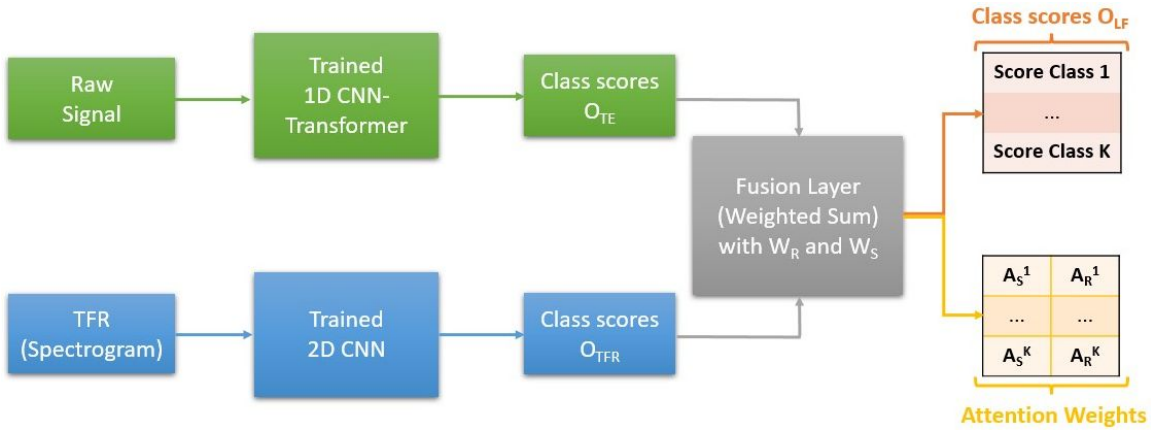

Figure 1: General pipeline of the proposed late fusion method. The green branch corresponds to the 1D-CNN-Transformer model extracting features from the raw signal. The blue branch corresponds to the 2D CNN model extracting features from the TFR. $\mathbf{W_R}$ and $\mathbf{W_S}$ are the raw signal and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

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 obtained 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 as $\mathbf{H_{TE}}$, can be combined with hidden features from other representations of the raw signal, or it can be fed to a specific classifier to do classification. If classification is done, we denote as $\mathbf{O_{TE}} \in \mathbb{R}^{K\times 1}$ 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's 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 feature from the TFR, we used a conventional 2D CNN architecture (each spectrogram is processed as an image). A summary of the used architecture can be find in figure3. 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, denotes as $\mathbf{H_{TFR}}$, can be combined with hidden feature from 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.

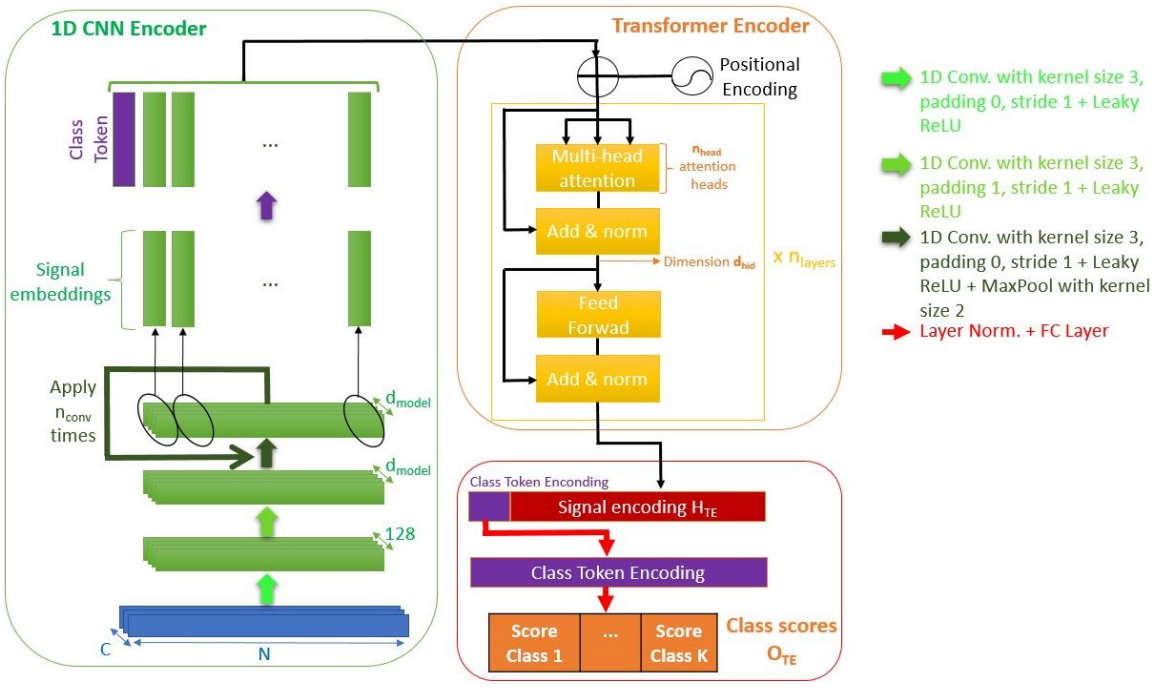

Figure 2: Hybrid 1D CNN-Transformer architecture

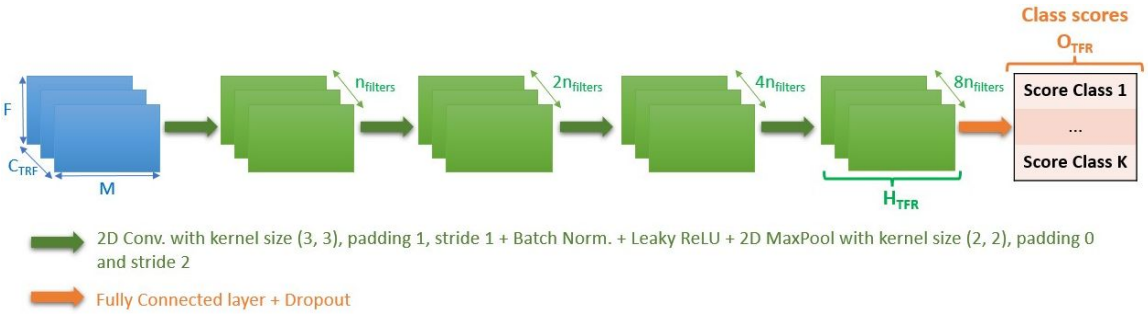

Figure 3: 2D CNN architecture used for classification using as input a time-frequency representation.

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's denote by $\mathbf{W_R} \in \mathbb{R}^{K\times 1}$ the attention weight vector associated with to 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 $\odot$ represent 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 finished, 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 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, as $\mathbf{H_{TE}}$ 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 give 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 by $\mathbf{H_{cat}}$ the concatenated feature, $\mathbf{H_{sum}}$ the summed feature and by $\mathbf{H_{att\_sum}}$ 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 indicating the global importance of each representation for the final decision of the model.

Finally, the obtained representation is passed through a FC layer of shape $64 \times K$ to do 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 electrocardigram (ECG) public 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: $8 - 10\ mm^3$.

Table 1 describe the number of samples per class and 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), an $n_{overlap} = 8$ (size of the overlap) and a Blackman window. 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 corresponding dataset split. Finally, the spectrograms of each 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 resampling at a frequency of 125 Hz. The PTB dataset focusing on Myocardial Infarction identification (two classes) and the MIT-BIH dataset focusing on Arrhythmia classification (5 classes). We used the standardized version of both datasets presented in (Kachuee et al., 2018)[1]. 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. Finally, the authors also proposed a training, validation and testing splitting which was also used in this paper. Tables 2 and 3 describe

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

Table 1: Number of samples per class in the HITS dataset. The unknown class correspond to HITS that could not be annotated.

Class	Number of samples
Artifact	403
Gaseous emboli	569
Solid emboli	569
Unknown	4

Table 2: Number of samples per class in the PTB dataset.

Class	Number of samples
Normal	10506
Abnormal	4046

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).

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 experiments 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 performances 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 4 presents the training parameters of the different models. All the models where trained using Cross Entropy (CE) loss, with class weights to handle the imbalanced classes. The class weights were computed using Scikit Learn (Pedregosa et al., 2011) and their

Table 3: Number of samples per class in the MIT-BIH dataset. Each of the classes regroup a set of abnormal hearbeats. 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

Table 4: 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 MIT-BIH				150
2D CNN	HITS	10^{-5}	10^{-5}	4	40
	PTB MIT-BIH	10^{-3}		16	30
Hybrid	HITS	10^{-2}	10^{-8}	16	15
	PTB	10^{-3}	10^{-2}		10
	MIT-BIH	3×10^{-4}	10^{-2}		

approach is inspired from (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 warmup steps.

All the experiments were repeated 10 times and the 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 5. 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 representations 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 0.5% in terms of accuracy. Thirdly, we can see that globally, using multiple representations allow to reduce the variability of the mean performances of the model, reducing in the best case by 0.45%. Moreover, for the MIT-BIH dataset, we get close performances to the

Table 5: 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	$\mathbf{89.33 \pm 2.77}$	$\mathbf{91.15 \pm 1.97}$	$\mathbf{93.39 \pm 1.74}$
PTB	MIF (Ahmad et al., 2021)	-	-	98.4
	MFF (Ahmad et al., 2021)	-	-	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	$\mathbf{99.29 \pm 0.21}$	$\mathbf{99.65 \pm 0.10}$	$\mathbf{99.71 \pm 0.08}$
MIT-BIH	MIF (Ahmad et al., 2021)	-	-	98.6
	MFF (Ahmad et al., 2021)	-	-	**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	$\mathbf{94.63 \pm 0.29}$	$\mathbf{91.28 \pm 0.54}$	98.37 ± 0.09

MIF approach (98.4% against 98.6%) of (Ahmad et al., 2021) but we are not able to reach the performances of MIF (it outperforms our method by 1.3%). However, in section 6 we discuss further about the relevance of the accuracy metric when dealing with imbalanced classes. Finally, tables 7, 8 and 9 show the final attention weights for each class and each representations. 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 with respect to other fusion method. 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 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 warmup steps. Additionally, we applied early stopping by selecting the model at the epoch with the maximum validation accuracy. All the experiments were repeated 10 times and the mean performances were compared

Table 6: 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	$\mathbf{94.63 \pm 0.29}$	$\mathbf{91.28 \pm 0.54}$	$\mathbf{98.37 \pm 0.09}$

using the Matthews correlation coefficient (MCC), the F1-Score and the accuracy measured on the test set.

Results are shown in table 6. 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. Secondly, we can see that, globally, the late fusion method reduces considerably the variability of the performances of the model (this is particularly true for the PTB and MIT-BIH datasets, where the variability can reduced by 0.97%). Thirdly, comparing with the results of experiment 1 (Table 5) we can notice that the three types of intermediate fusion does not improve the performances 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 a MCC degradation of 1.3% the PTB dataset. Finally, we conclude that the performance of the three intermediate fusion methods are very close and none of them competes with the late fusion approach.

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 of using our proposed method to improve the classification performances of a model in three different medical datasets. Our proposed method takes advantage of the complementary of both representations, the raw signal focusing on the temporal context and the amplitude information, whereas the spectrogram focusing 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 it showed the same behaviour 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 from the randomness of the training procedure.

What is more, 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 and F1-score because we are dealing with higlhy imbalanced datasets (specially the PTB and MIT-BIH datasets). By the same token, we can see that the performances on the HITS dataset are smaller than the ones obtained on PTB or MIT-BIH. This can be explained by two main reasons: the dataset size, 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 hyper-parameters that has to be optimized. Thirdly, the multiple features late fusion model is heavier in terms of memory than single feature models as we increase the number of parameters.

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 lead 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 therms 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

Table 7: Attention weights median values and mean absolute deviations for the late fusion model on the HITS dataset

Class	Spectrogram	Raw Signal
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

the HITS dataset, the best performing feature is the spectrogram (contrary to PTB and MIT-BIH) which is the one having the greater variability. Second, as the attention weights of table 7 show it, the final decision of the hybrid model are 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.

This last point illustrates the interest of the attention weights for interpetability purposes. Indeed, our method offers interpretable attention weights for each representation and for each class as showed in tables 7, 8 and 9. This can give interesting insight 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 discriminate enough to classify the case. However, in some cases, the expert can hesitate and use the raw signal to remove the 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 advantages with respect to the other presented fusion approaches: it is easier to optimize. Indeed, we just need to optimize each single feature model independently and then we finetune the attention weights which is not a difficult task. For the intermediate fusion methods we add FC layers which add extra parameters and hyper-parameters, 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 do classification with raw signals and TFRs. Instead of choosing one fixed initial representation of the signal, our method proposed to exploit two complementary representation: the raw signal (temporal information) and the spectrogram (spectral information). We pass these two representa-

Table 8: Attention weights median values and mean absolute deviations for the late fusion model on the PTB dataset

Class	**Spectrogram**	**Raw Signal**
Normal	0.49 ± 0.12	0.51 ± 0.12
Abnormal	0.18 ± 0.10	0.82 ± 0.10

Table 9: Attention weights median values and mean absolute deviations for the late fusion model on the MIT-BIH dataset

Class	**Spectrogram**	**Raw Signal**
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

tions 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. This weights attribute an importance of each representation for the final classification score of each class. Extensive experiments on 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 theLABEX CELYA (ANR-10-LABX-0060) and PRIMES (ANR-11-LABX-0063) of Universite de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

Asad Abdi, Siti Mariyam Shamsuddin, and Jalil Piran. Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing Management*, 56:1245–1259, 02 2019. doi: 10.1016/j.ipm.2019.02.018.

Asad Abdi, Shafaatunnur Hasan, Siti Mariyam Shamsuddin, Norisma Idris, and Jalil Piran. A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion. *Knowledge-Based Systems*, 213:106658, 2021.

ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.106658. URL https://www.sciencedirect.com/science/article/pii/S0950705120307875.

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.

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.

Himanshu Chaurasiya. Time-frequency representations: Spectrogram, cochleogram and correlogram. *Procedia Computer Science*, 167:1901–1910, 2020.

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.

Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968, 2014. doi: 10.1109/ICASSP.2014.6854950.

Eric Donkor. Stroke in the 2 1 s t century: A snapshot of the burden, epidemiology, and quality of life. *Stroke Research and Treatment*, 2018:1–10, 11 2018. doi: 10.1155/2018/3238165.

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.

Ömer Faruk Ertugrul, Emrullah Acar, Erdoğan Aldemir, and Abdulkerim Öztekin. Automatic diagnosis of cardiovascular disorders by sub images of the ecg signal using multi-feature extraction methods and randomized neural network. *Biomed. Signal Process. Control.*, 64:102260, 2021.

Xin Feng, Qiang Feng, Shaohui Li, Xingwei Hou, and Shugui Liu. A deep-learning-based oil-well-testing stage interpretation model integrating multi-feature extraction methods. *Energies*, 13(8), 2020. ISSN 1996-1073. doi: 10.3390/en13082042. URL https://www.mdpi.com/1996-1073/13/8/2042.

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.

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.

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.

Mitesh M. Khapra, A Kumaran, and Pushpak Bhattacharyya. Everybody loves a rich cousin: An empirical study of transliteration through bridge languages. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 420–428, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/N10-1065.

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.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

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.

Zhang-Meng Liu. Multi-feature fusion for specific emitter identification via deep ensemble learning. *Digital Signal Processing*, 110:102939, 2021. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2020.102939. URL https://www.sciencedirect.com/science/article/pii/S1051200420302840.

Shihan Mao, Yuhua Li, You Ma, Baohua Zhang, Jun Zhou, and Kai Wang. Automatic cucumber recognition algorithm for harvesting robots in the natural environment using deep learning and multi-feature fusion. *Computers and Electronics in Agriculture*, 170, 02 2020. doi: 10.1016/j.compag.2020.105254.

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.

Juan D. S. Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro Lameiras Koerich. Multimodal fusion with deep neural networks for audio-video emotion recognition. *ArXiv*, abs/1907.03196, 2019.

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.

Hyunsin Park and Chang D. Yoo. Cnn-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification. *IEEE Signal Processing Letters*, 27:411–415, 2020. doi: 10.1109/LSP.2020.2975422.

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, Benjamin Blankertz, and Klaus Obermayer. A new method for tracking modulations in tonal music in audio data format. volume 6, pages 270–275 vol.6, 02 2000. ISBN 0-7695-0619-4. doi: 10.1109/IJCNN.2000.859408.

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.

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.

Shahriar Shariat and Vladimir Pavlovic. Isotonic cca for sequence alignment and activity recognition. In *2011 International Conference on Computer Vision*, pages 2572–2578, 2011. doi: 10.1109/ICCV.2011.6126545.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/2d6cc4b2d139a53512fb8cbb3086ae2e-Paper.pdf.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. doi: 10.1162/tacl_a_00177. URL https://aclanthology.org/Q14-1017.

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.

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5005–5013, 2016. doi: 10.1109/CVPR.2016.541.

Lizhe Wang, Jiabin Zhang, Peng Liu, Kim-Kwang Raymond Choo, and Fang Huang. Spectral–spatial multi-feature-based deep learning for hyperspectral remote sensing image classification. *Soft Computing*, 21, 01 2017. doi: 10.1007/s00500-016-2246-3.

Jael Sanyanda Wekesa, Jun Meng, and Yushi Luan. Multi-feature fusion for deep learning to predict plant lncrna-protein interaction. *Genomics*, 112(5):2928–2936, 2020. ISSN 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2020.05.005. URL https://www.sciencedirect.com/science/article/pii/S088875431931016X.

Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2012.03.001. URL https://www.sciencedirect.com/science/article/pii/S0262885612000285. Affect Analysis In Continuous Input.

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.

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. Interpretability of the Attention Weights using Integrated Gradients

We aim at interpreting the attention weights using the *Integrated Gradients* (IG) [REF] method.

Let $f : \mathbb{R}^n \to \mathbb{R}$. The integrated gradient of an input feature $x_i \in \mathbb{R}$ of an input vector $x \in \mathbb{R}$ with respect to a baseline $x' \in \mathbb{R}^n$ is defined by:

$$\phi_i(f, x, x') = (x_i - x'_i) \times \int_0^1 \frac{\partial f(x' + \alpha \times (x - x')}{\partial x_i} \, d\alpha \tag{7}$$

In our case, we want to interpret the final classification of the late fusion hybrid model with respect to the classification outputs of each single feature model. This will allow us to better understand the behaviour of the learned attention weights. Let's define the classification function to study as follows:

$$\begin{aligned} f \colon \mathbb{R}^{C \times 2} &\to \mathbb{R}^C \\ X &\mapsto X \times B^T \end{aligned}$$

with:

- C is the number of classes of the classification problem.
- $X = \begin{bmatrix} x_{TFR}^1 & x_{TE}^1 \\ ... & ... \\ x_{TFR}^C & x_{TE}^C \end{bmatrix}$
- $B = \begin{bmatrix} \beta_{TFR}^1 & \beta_{TE}^1 \\ ... & ... \\ \beta_{TFR}^C & \beta_{TE}^C \end{bmatrix}$

Let's now denote the i^{th} component of the output of f as follows:

$$\begin{aligned} f_i \colon \mathbb{R}^{C\times 2} &\to \mathbb{R} \\ X &\mapsto \beta_{TFR}^i \times X_{i,1} + \beta_{TE}^i \times X_{i,2} \end{aligned}$$

Now, we can apply IG to each component of the output of f to determine the importance of the classification output of a single feature model to the final decision of the hybrid classification model. For all $i \in [1, C]$, for all $X = [x_{TFR}, x_{TE}], X' = [x'_{TFR}, x'_{TE}] \in \mathbb{R}^{C\times 2}$ with $x_{TFR}, x'_{TFR}, x_{TE}, x'_{TE} \in \mathbb{R}^C$:

$$\begin{aligned} \phi_i^{TFR}(f_i, X, X') &= (x_{TFR}^i - {x_{TFR}^i}') \times \int_0^1 \beta_{TFR}^i \times \alpha \, d\alpha \\ &= \beta_{TFR}^i \times (x_{TFR}^i - {x_{TFR}^i}') \times \int_0^1 \alpha \, d\alpha \\ &= \frac{\beta_{TFR}^i}{2} \times (x_{TFR}^i - {x_{TFR}^i}') \end{aligned} \tag{8}$$

$\phi_i^{TFR}(f_i, x_i, x'_i)$ represents the importance of the prediction of the class i from the TFR model for the final prediction of class i of the hybrid model. By the same token, the prediction of the class i from the raw signal model for the final prediction of class i of the hybrid model is given by:

$$\phi_i^{TE}(f_i, X, X') = \frac{\beta_{TE}^i}{2} \times (x_{TE}^i - {x_{TE}^i}') \tag{9}$$

Moreover, we can also compute the importance of the prediction of the class i from the TFR/TE model for the final prediction of class $j \neq i$ of the hybrid model:

$$\phi_i^{TFR}(f_j, X, X') = (x_{TFR}^i - {x_{TFR}^i}') \times \int_0^1 0 \, d\alpha = 0 \tag{10}$$

By the same token, for all $i, j \in [1, C]$ such that $i \neq j$:

$$\phi_i^{TE}(f_j, X, X') = 0 \tag{11}$$

Finally, as it is usually done with IG, we are going to take as baselines, X', the vectors/matrices composed of only 0 elements. The final form of the importance of the prediction of the class i from the TFR/TE model for the final prediction of class is given by:

$$\phi_i^{TFR/TE}(f_j, X, X') = \begin{cases} \frac{\beta^i_{TFR/TE} \times x^i_{TFR/TE}}{2} \text{ if } i = j \\ 0 \text{ else} \end{cases}$$

This results are validates experimentally using the library Captum. **WARNING:** The results were validated experimentally but we got a factor 2 in the experiments.

This shows that our attention weights can also be interpreted as the contribution of the classification output of a single model to the classification of the final hybrid model.

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

Table 10: 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 embolu	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 11: 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 embolu	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