Information Fusion 100 (2023) 101955

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series

Duarte Folgado [a,b,*,1], Marília Barandas [a,b,1], Lorenzo Famiglini [c], Ricardo Santos [a,b], Federico Cabitza [c,d], Hugo Gamboa [a,b]

[a] Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, Porto, 4200-135, Portugal
[b] LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), NOVA School of Science and Technology, Campus de Caparica, 2829-516, Portugal
[c] Department of Informatics, Systemics and Communication, University of Milano-Bicocca, Viale Sarca 336, Milan, 20126, Italy
[d] IRCCS Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi 4, Milan, 20161, Italy

ARTICLE INFO

Keywords:
Explainable AI
Uncertainty quantification
Multimodal
Complexity
SHAP
Feature-based explanations

ABSTRACT

Feature importance evaluation is one of the prevalent approaches to interpreting Machine Learning (ML) models. A drawback of using these methods for high-dimensional datasets is that they often lead to high-dimensional explanation output that hinders human analysis. This is especially true for explaining multimodal ML models, where the problem's complexity is further exacerbated by the inclusion of multiple data modalities and an increase in the overall number of features. This work proposes a novel approach to lower the complexity of feature-based explanations. The proposed approach is based on uncertainty quantification techniques, allowing for a principled way of reducing the number of modalities required to explain the model's predictions. We evaluated our method in three multimodal datasets comprising physiological time series. Results show that the proposed method can reduce the complexity of the explanations while maintaining a high level of accuracy in the predictions. This study illustrates an innovative example of the intersection between the disciplines of uncertainty quantification and explainable artificial intelligence.

1. Introduction

Interpretability is a crucial aspect of Machine Learning (ML) systems, as it enables the provision of predictions and explanations of their outputs. By facilitating an enhanced understanding of the rationale behind the prediction, interpretable machine learning systems play a vital role in fostering trust and safety. This is achieved by providing insights into the prediction's reasonableness, enabling users to identify any areas of concern or potential risks [1].

Feature importance evaluation is one of the prevalent approaches for interpreting black-box ML models. It has been previously used across several domains, including healthcare [2–7], human activity recognition [8], credit risk management [9], and materials sciences and chemistry [10]. Evaluating feature importance consists of visualizing the importance ranking of each feature and measuring how much a given feature contributes to the prediction. Despite the versatility of feature importance approaches, they have occasionally limited usability. Molnar et al. [11] emphasize this shortcoming by identifying a pitfall in using these methods for high-dimensional datasets that often lead to an overwhelming and high-dimensional explanation output that hinders human analysis.

This issue becomes particularly relevant in explaining multimodal machine learning models, where the complexity of the problem is aggravated by an increase in the total number of features and the variety of data modalities being considered. An example scenario is wearable data, where multiple sensors are worn across multiple body regions and retrieve information from multiple data modalities. We define *modalities* as distinct types or categories of data, such as motion, heart rate variability, and respiration. Each modality consists of one or more *features*, which are quantitative measures derived from data representing the statistical, temporal, and spectral properties of each modality. A single model that learns from these high-dimensional datasets is often particularly complex and might attempt to model dozens of features from a broad spectrum of data modalities.

Multimodal learning attempts to model the combination of different modalities of data. However, it is often found that improved

* Corresponding author at: LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), NOVA School of Science and Technology, Campus de Caparica, 2829-516, Portugal.
E-mail address: d.folgado@campus.fct.unl.pt (D. Folgado).
[1] These authors contributed equally.

https://doi.org/10.1016/j.inffus.2023.101955
Received 17 April 2023; Received in revised form 28 June 2023; Accepted 26 July 2023
Available online 29 July 2023

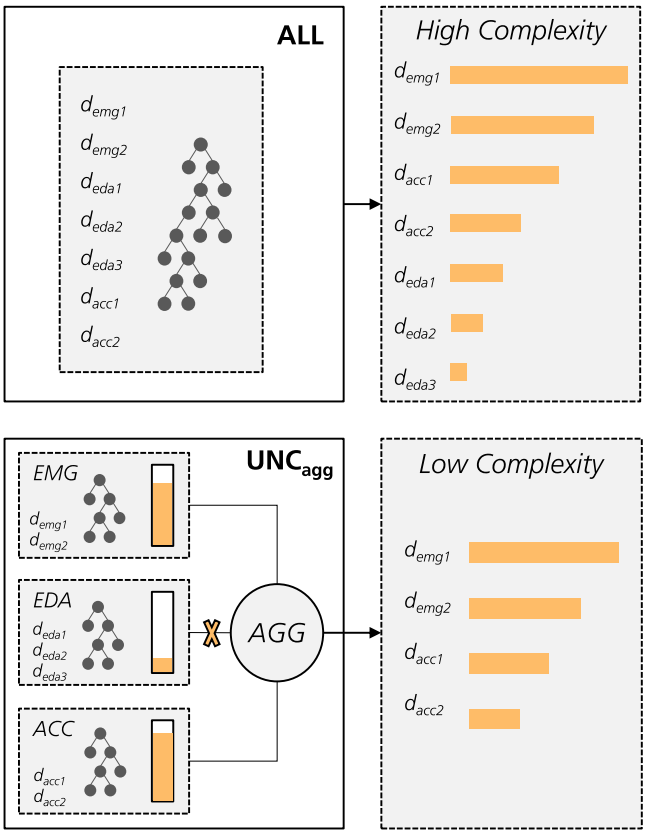


Fig. 1. Schematic representation of our framework. The upper panel shows an early fusion model that learns from different modalities (ALL), resulting in a complex high-dimensional explanation output that hinders human analysis. In the lower panel, we propose a new approach that uses uncertainty quantification for model late fusion by rejecting the most uncertain models (UNC_{agg}). The vertical orange bars represent the model's confidence levels. By reducing the number of modalities and features, our approach results in explanations with lower complexity.

performance can be obtained by combining multiple models instead of relying on a single model [12,13]. Multimodal data fusion is the task of joining information from two or more modalities to perform the prediction. The most straightforward approach to combining models is calculating the mode from the predictions of a group of models. Such a methodology is justified from a frequentist viewpoint by weighing bias and variance. However, this combination process assumes that all models perform equally well regardless of their reliability or predictive uncertainty. Alternative late fusion methods can incorporate dynamic weights based on Uncertainty Quantification (UQ) for the aggregation process. Dynamic weighting enables the ensemble model to consider the uncertainty level of each individual model prediction, assigning weights that are inversely proportional to the prediction's uncertainty.

Insights from social sciences [14] support simple model explanations that involve the minimum number of features. Therefore, it is important to research methodologies for reducing the explanation complexity, particularly in multimodal scenarios.

In this paper, we propose a novel approach to lower the explanation complexity of multimodal data using uncertainty quantification. Fig. 1 outlines the intuition of the proposed approach. A baseline model uses all the data modalities of the dataset with early fusion. We use UQ to aggregate specialized models for each modality by accounting for their respective uncertainty in predicting for a given sample. This process allows for discarding less confident models and using a subset of modalities for making predictions. By reducing the number of modalities taken into account, the complexity of the explanation is also minimized.

Contributions. We show empirical evidence that late model fusion using UQ methods helps reduce the complexity of explanations provided by feature importance scores. To support our claim, we provide an in-depth study on systematically evaluating different model late fusion methods using UQ and introduce a novel measure for quantifying explanation complexity in multimodal datasets. In a more broad view of research on ML, this study also introduces an innovative example of how we can leverage the intersection between the disciplines of UQ and model explainability towards the topic of responsible Artificial Intelligence (AI).

The rest of this paper is organized as follows. Section 2 provides an overview of the related work in feature-based explanations and model fusion approaches. Section 3 presents our main contributions and proposed approaches to lowering explanation complexity via model combination using UQ. Section 4 describes our experiments' datasets, pipelines, and results. Section 5 summarizes our findings and discusses the implications of our results. Finally, Section 6 concludes the paper, suggests directions for future work, and evaluates the broad impact of this research.

2. Background

2.1. Feature-based explanations

Research on improving the interpretability of black-box ML models through post-hoc explanations has attracted considerable attention over the last few years. Feature-based explanations are popular among practitioners who want to understand their model better to ensure its adequate behavior when deploying in real-world applications. These techniques often involve visualizing the importance ranking of each feature and how the feature values affect the model's prediction. Formally, feature-based methods assign a scalar attribution value, sometimes called "relevance" or "contribution" to each input sample's input feature. The goal is to determine the contribution $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_d] \in \mathbb{R}^d$ of each input feature x_i to the output $\hat{y}_i$.

Several approaches are available to measure each feature's relevance across different data types. Gradient-based methods compute the gradient of the model's output to its inputs and use those values to represent in salience maps [15]. Perturbation-based methods modify or remove parts of the input and measure the impact on the model's output [16]. In the latter setting, Local Interpretable Model-Agnostic Explanations (LIME) is a well-known technique that approximates the model locally with a simpler surrogate model for which several perturbations are performed [17]. Another well-known method is SHapley Additive exPlanations (SHAP), which translates the Shapley values from cooperative game theory into the context of ML [18].

The evaluation of the explanation quality is a non-trivial task due to its subjective nature. Standardizing such evaluation is an open research topic that has received significant attention [19–21]. Insights from social sciences point out that interpretability has properties of clarity and parsimony. Clarity implies that the explanation is unambiguous, while parsimony means that the explanation is presented in a simple and compact form [22]. Lombrozo [14] argues that good explanations are simple and broad. The author's observation is supported by user research studies involving academics across several disciplines that identify consistency in their judgment regarding what constitutes good explanations. Appeals to simplicity were ubiquitous, and some participants also emphasized the significance of generality or comprehensiveness.

The literature often uses model complexity as a proxy for explainability complexity. In addition to the number of features, some model-specific metrics are used, such as the number of decision tree rules, tree depth, and non-zero coefficients in linear models [23–25]. This approach assumes that the more parameters a model has, the more complex it is. Another popular approach is to use information criteria such as the Akaike information criterion or the Bayesian information

criterion [26]. These criteria provide means to compare the relative complexity of different models while also considering their goodness of fit. L1 regularization (Lasso) is a popular method for feature selection, as it shrinks the less important feature's coefficient to zero, thus, removing some features altogether. Prior studies have employed a technique of modifying the method by increasing the emphasis on minimizing particular modalities while assigning differential weights to features based on their modalities or aggregating features belonging to the same modality and subjecting the group to a penalty [27–30].

Bhatt et al. introduced in [31] three criteria for evaluating feature-based explanations: sensitivity, faithfulness, and complexity. Sensitivity measures the change in the explanation after perturbation in the model input; faithfulness concerns the capacity of an explanation method to select the truly relevant features; and complexity concerns the extent of the simplicity of the explanation. The authors present a desideratum for good explanations: low sensitivity, high faithfulness, and low complexity. Batterman and Rice studied the complexity of explanations in [32] and argued for minimal model explanations that contain only relevant and representative features. In fact, humans cannot process a high volume of information at once, therefore, explanations should have reduced complexity (i.e., use few features). To examine the influence of explanation intricacy on users' understanding, Lage et al. [33] investigated the impact of explanation length and complexity on response time, accuracy, and the subjective satisfaction of users. Their findings indicated that heightened explanation complexity led to a decrease in subjective user satisfaction.

2.2. Model fusion and weighting schemes

In particularly complex classification problems, performance can be improved by combining multiple models instead of just using one. It has been observed that although one model would yield the best performance for a given classification task, the sets of observations misclassified by the different models would not necessarily overlap. This suggested that different models potentially offered complementary information about the observations to be classified, which could be harnessed to improve the performance of the selected model [12]. Ensembles generally require heterogeneity of predictions to be successful, regardless of the combination rule. This can be achieved through different feature sets or parameter settings for identical learning models or through different learning algorithms using the same features. The key is to avoid identical erroneous decisions on the same observation instances so that the individual classifiers provide complementary information. This approach appears in the literature under several names, such as multi-classifiers combination, multi-classifiers fusion, a mixture of experts, and ensemble-based classification systems, among others [13].

Despite the idea of combining models is not new, an interesting topic discussed in the research community is how to find the optimal combination rule for a given task. There are several late fusion rules to train and combine different models. Some rules address model combinations using voting mechanisms based on individual classifiers' predictions, while others use aggregation techniques based on the classifiers' class probabilities. The former is commonly called hard voting (also known as majority voting), whereas the latter is soft voting.

In the literature, using weights in the model combination process is common since models often exhibit varying performance levels. Some studies, such as those by [34] and [35], leverage weights based on models' reliability scores, outperforming baseline methods with equal weights. However, dynamic weighting schemes tend to yield better results than fixed weights [13]. This superior performance can be attributed to the dynamic combination's ability to update weights assigned to individual classifiers before making the final decision. For example, Poh et al. [36] proposed a quality-based combination approach for multimodal biometrics. The underlying concept is that quality issues affecting one modality (e.g., signal noise) often do not impact other modalities. Consequently, the proposed combination method assigns higher weights to more reliable classifiers under specific conditions.

Different combination strategies exist, but the ensemble's individual classifiers' predictive uncertainty is seldom considered [37]. In the following section, we will discuss different aggregation strategies and present various measures of UQ that we propose to use as weights for the model aggregation methods.

2.2.1. Aggregation strategies

Let us consider a standard setting of supervised learning with a finite training dataset, $D = \{(x_i, y_i)\}_i^N \subset \mathcal{X} \times \mathcal{Y}$, with N samples, composed of pairs of input instances x and outcomes y, where $\mathcal{X}$ is an instance space, $\mathcal{Y}$ the set of outcomes that can be associated with an instance. Suppose a hypothesis space $\mathcal{H}$ composed by a finite ensemble of M hypothesis, where a hypothesis h maps instances x to outcomes y. An individual model can be seen as a hypothesis of the ensemble.

One of the most common forms of aggregation is majority voting. As the name suggests, the predicted class label is obtained by considering the vote of each classifier with equal importance, i.e., the final prediction is the most frequently predicted class label. This method offers the benefit of directly handling the outputs of individual classifiers without the need for probabilistic modeling. However, it assumes that all classifiers perform equally well regardless of their reliability or predictive uncertainty [38]. In general, the individual models' performances are not similar, so it is reasonable to assign higher weights to the decision made by the more accurate classifiers using a weighted majority voting defined as:

$$agg_{vote}(x) = \arg\max_{y \in \mathcal{Y}} \sum_{h \in \mathcal{H}} w_h * [\![\hat{y}_h = y]\!] \tag{1}$$

where $w_h \in \mathbb{R}^+$ denotes the weight associated with hypothesis h and $[\![\hat{y}_h = y]\!]$ is an indicator function that takes the value 1 if the expression is true, and 0 otherwise. We recover standard majority voting if $w_h = 1$ for all M hypotheses in the ensemble.

Instead of using the predicted class label, aggregation based on the class probabilities of each classifier can be made. One of the most straightforward methods in this soft-level aggregation is the average or sum of class probabilities. Although these soft aggregation methods consider more information in the combination process, they require different models to approximate the same function. Otherwise, the predictions are incomparable, and averaging is not a meaningful operation [39]. Moreover, calibrating individual classifiers' probabilities can be challenging in the combination process.

These aggregations can also be turned into a weighted version. As an example, the sum rule quantifies the likelihood of a hypothesis by combining the class probabilities generated by the individual ensemble members using a weighted sum rule defined as:

$$agg_{sum}(x) = \arg\max_{y \in \mathcal{Y}} \sum_{h \in \mathcal{H}} w_h * p(y|x, h) \tag{2}$$

where $p(y|x, h)$ is the probability of outcome y given x predicted by hypothesis h.

This definition can be generalized to other combination rules, such as the average, product, maximum, minimum, and median.

2.2.2. Uncertainty quantification measures

In recent years, researchers have shown an increased interest in estimating uncertainty in ML. The most common method for estimating the uncertainty of a prediction, known as predictive uncertainty, involves separately modeling aleatoric and epistemic uncertainty. Aleatoric uncertainty refers to the notion of randomness arising from data complexity, multi-modality, and noise. On the other hand, epistemic uncertainty is caused by a lack of knowledge of the underlying process being modeled, either due to the uncertainty associated with the model or the lack of data [40].

Regarding UQ measures, a straightforward way of quantifying uncertainty is using the output of the classification task that represents

the class probabilities. For a given observation x, the probability of the predicted class, or maximum probability, can be obtained by the following equation:

$$p(\hat{y}|x) = \max_{y \in \mathcal{Y}} p(y|x) \tag{3}$$

Additionally, the entropy of the predictive posterior, modeled by the (Shannon) entropy, is the most well-known measure of uncertainty of a single probability distribution. For discrete class labels is given by Eq. (4):

$$H[p(y|x)] = -\sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \tag{4}$$

Both the maximum probability and entropy of the predictive posterior distribution can be seen as measures of aleatoric uncertainty in predictions.

Assuming a Bayesian perspective, where the predictive posterior distribution is approximated using ensemble techniques, the decomposition of aleatoric and epistemic uncertainty can be obtained using the proposed approach of Depeweg et al. [41]. The authors proposed to measure the total uncertainty (Eq. (5)) in terms of the entropy of the predictive posterior distribution, the aleatoric uncertainty (Eq. (6)) in terms of the expectation of entropy concerning the posterior probability, and the epistemic uncertainty in terms of the mutual information between hypotheses and outcomes, approximated by the difference between total uncertainty and aleatoric uncertainty $u_e(x) = u_t(x) - u_a(x)$.

$$u_t(x) := -\sum_{y \in \mathcal{Y}} \left(\frac{1}{M} \sum_{h \in \mathcal{H}} p(y|x,h) \right) \log_2 \left(\frac{1}{M} \sum_{h \in \mathcal{H}} p(y|x,h) \right) \tag{5}$$

$$u_a(x) := -\frac{1}{M} \sum_{h \in \mathcal{H}} \sum_{y \in \mathcal{Y}} p(y|x,h) \log_2 p(y|x,h) \tag{6}$$

Additionally, the variation ratio is another measure of epistemic uncertainty. It measures the variability of predictions by computing the fraction of samples with the correct output. This heuristic is a measure of the dispersion of the predictions around its mode [42] given by:

$$vr(x) = 1 - \frac{\sum_{h \in \mathcal{H}} [\![\hat{y}_h = \hat{y}]\!]}{M} \tag{7}$$

where $\hat{y}$ corresponds to the majority class obtained using $agg_{vote}(x)$ from Eq. (1).

3. From uncertainty quantification to simpler explanations

This section describes the methods of the main contributions of this work. While various methods can be used to describe the feature importance of combined models, we have opted to use SHAP and will start by addressing its formal definition.

3.1. Shapley values for feature importance

The Shapley values are a solution concept in cooperative game theory. They denote a player's marginal contribution to a coalitional game's payoff. Let T be a set of players and let $v : 2^T \to \mathbb{R}$ be the characteristic function, where $v(S)$ denotes the contribution of the players in $S \subseteq T$. The Shapley value of player j's contribution (i.e., averaging player j's marginal contributions to all possible subsets S) is:

$$\phi_j(v) = \frac{1}{|T|} \sum_{S \subseteq T \setminus \{j\}} \binom{|T|-1}{|S|}^{-1} (v(S \cup \{j\}) - v(S)). \tag{8}$$

In feature-based explanations, the problem is formulated similarly, and the game's payoff is the model's output $\hat{y} = f(x)$, the players are the d features of x, and the ϕ_j represent the feature contribution to the outcome $f(x)$.

SHAP calculates Shapley value explanations with an additive feature attribution method:

$$g\left(z'\right) = \phi_0 + \sum_{j=1}^{Z} \phi_j z'_j \,, \tag{9}$$

where g is the explanation model, $z' \in \{0,1\}^Z$ is the coalition vector, Z is the maximum coalition size, and $\phi_j \in \mathbb{R}$ is the feature attribution for feature j, i.e., the Shapley value.

3.2. Measuring explanation complexity from feature importance

In this section we introduce our proposed approach to measure the explanation complexity of a single instance, i.e., local explanation, using the feature importance.

The rationale for our approach resides in the considerations associated with multimodal problems. In this type of problem, features are associated with different modalities. We argue that a simpler explanation uses the smallest possible set of modalities and features. Therefore, let us consider a dataset D composed of a set of R features and M disjoint modalities. To explain the instance x_i, only a subset of $r \subseteq R$ features and $m \subseteq M$ models are required. We define complexity as the fraction of features and modalities required to explain a given instance in relation to the total number of features and modalities.

Definition 1. Given an explanation function $\boldsymbol{g}$ that depends on the subset of r features and the subset of m modalities to explain the instance x_i, the complexity of $\boldsymbol{g}(r,m)$ at x_i is:

$$c(g; x_i) = \frac{1}{2} \left(\frac{|r|}{|R|} + \frac{|m|}{|M|} \right) \,, \tag{10}$$

where $|\cdot|$ is the cardinality of a set and $c \in [0,1]$.

We provide a practical note with an illustrative example in Appendix A.

A complex explanation uses $|R|$ features and $|M|$ modalities. Although the explanation is probably faithful to the model, it is difficult for the user to understand the relationships between the high number of features and different modalities contributing to a given prediction.

The following heuristic might be followed to find the subset r. Suppose the SHAP values are rescaled from the log odds to the probability space. In that case, their sum equals the difference between the posterior probability, $p(y|x_i)$, and the expected base value $\mathbb{E}[f(x_i)]$. The base value is generally the average of the outcome variable in the training set. We define the minimum number of relevant features as the minimum subset from which the sum of Shapley values, ordered by their absolute value, $\bar{\phi}_j$, equals or surpasses the difference between the posterior probability and the base value.

Definition 2. Given the set of R features and their Shapley values ordered by their absolute value, $\bar{\phi}_j$, the minimum relevant feature subset, r, to explain the instance x_i, is the one that satisfies the following:

$$\min \left\{ r \subseteq R : \sum_{j=1}^{|R|} \bar{\phi}_j \right\} \geq \left| p(y|x_i) - \mathbb{E}[f(x_i)] \right| . \tag{11}$$

Based upon the axiomatic assumptions of SHAP, local explanations from a large number of samples can be combined to have global model insights while preserving local faithfulness to the original model [43,44], which holds true for structured [45], and time series datasets [46]. Usually, this is accomplished by averaging the absolute SHAP value of each feature across all samples. In the same spirit, the global explanation complexity for a given model can be calculated by averaging all the local explanation complexities.

Definition 3. Given the set of c explanation complexities of N samples, the global explanation complexity of $\boldsymbol{g}$ is:

$$C(c) = \frac{1}{N} \sum_{i=1}^{N} c_i. \tag{12}$$

3.3. Uncertainty-weighted late model fusion strategies

We propose using both soft and hard aggregation strategies, as described in Section 2.2.1, for the task of model combination, weighted at two main levels:

- **Model-based**: The predictions of each individual model are weighted by the model's classification performance. This approach is grounded in the principle that the more accurate models should be given greater weight than the less accurate ones, as they are more likely to provide reliable predictions.
- **Instance-based**: The predictions of each individual model are dynamically weighted using measures of uncertainty (see Section 2.2.2) for a given instance. Dynamic weighting enables the ensemble model to consider the prediction's confidence from the individual models. Although one model may generally be more accurate than the others, it does not necessarily mean it will always be more certain in its decisions. In some cases, the most accurate model may exhibit higher uncertainty in its predictions than the less accurate model.

Uncertainty measures can be used in both soft and hard aggregation strategies. However, given that aleatoric measures are derived from class probabilities and soft voting relies on probability values, we have opted to exclusively use epistemic uncertainty measures for soft voting. Conversely, for hard voting, we have employed both aleatoric and epistemic uncertainties.

Additionally, we propose a combination strategy based on classification with a rejection option for individual models. For instance-based weights, a model with high uncertainty in a given observation will correspond to a low weight in the combination process. Therefore, its contribution to the decision process will be minimal. However, in the case of model-based weights, the abstaining capabilities of individual models can be advantageous for the combination process. Thus, we proposed to use both aleatoric and epistemic uncertainty as rejection measures for hard voting strategies and only epistemic uncertainty for soft voting. Defining a suitable uncertainty rejection threshold is a challenging task that goes beyond the scope of this work. Therefore, rather than attempting to learn an optimal rejection threshold, we have established the rejection threshold by considering a percentage of the maximum theoretical value for each uncertainty measure. Since all the uncertainty measures used in this study are upper and lower bounded, we have computed the maximum theoretical value for each measure and set the rejection threshold at 90% of that value.

Table 1 summarizes the late model fusion strategies studied in the experimental analysis.

Table 1
Summary of combination strategies weighted by uncertainty measures. AU: Aleatoric Uncertainty, EU: Epistemic Uncertainty, TU: Total Uncertainty, MR: Model Reliability.

Aggregation	Uncertainty	Metric	Abbreviation
Hard voting	Aleatoric	$p(\hat{y}\|x)$	$AU_{\max}$
		$H[p(y\|x)]$	AU_{entropy}
		$u_a(x)$	AU_{bayes}
	Epistemic	$vr(x)$	EU_{vr}
		$u_e(x)$	EU_{bayes}
	Total	*	TU_*
	Model Reliability	Performance	MR
		Performance (AU and EU based rejection)	MR_{rej}
Soft voting	Epistemic	$vr(x)$	EU_{vr}
		$u_e(x)$	EU_{bayes}
	Model Reliability	Performance	MR
		Performance (EU based rejection)	MR_{rej}

*Represents any combination of aleatoric and epistemic uncertainty measures.

3.4. Lowering explanation complexity via model combination using uncertainty

We hypothesize that reducing the number of modalities and features used in the explanation can simplify the model's explanation complexity. When modeling multimodal data, one can either use a single model for all available features and modalities or use separate models for each modality and combine their strengths and consider them using the late model fusion strategies previously defined in Section 2.2.1.

For the model aggregation strategy, while the individual models might produce accurate predictions in general, in certain circumstances, this may not be the case. For example, certain regions of the feature space could exist where some models concerning specific modalities struggle to differentiate among the different classes. Using the combination strategy with a rejecting option proposed in Section 3.3, these models would abstain from making a prediction when they are likely to misclassify. Another advantage of this approach is that, since the model rejects uncertainty-based modalities, it decreases the number of modalities and features required to explain a particular instance, thus lowering the complexity of the explanation.

Different modalities might have different sizes of feature sets. However, all modalities are treated equally concerning complexity, irrespective of their associated feature counts, since it is not trivial to weigh modalities according to their number of features. Without contextual information about the modalities and features, it is difficult to attribute variable complexity weights to modalities according to the number of features. In fact, inherent high complexity might exist towards a given modality regardless of the number of features. Different levels of domain knowledge among subjects analyzing the explanation lead to variability and subjectivity w.r.t. to the perception of complexity for a given modality.

For each instance, we decrease the number of modalities and features to explain by: (1) rejecting the models with high uncertainty and (2) using only the individual models that are in agreement.

In order to calculate the Shapley values for the proposed approach, the posterior probability of the late model fusion strategy must be calculated. The computation varies depending on the aggregation strategy employed. In hard voting strategies, the class probability is determined by the fraction of votes of each modality, whereas in soft voting strategies, the class probabilities are computed as the mean predicted class probabilities of each model in the ensemble. The Shapley values are calculated for each feature within each modality separately.

The following sections show that this approach reduces the explainability complexity without compromising the overall model performance.

4. Experimental evaluation

4.1. Datasets

We tested our proposed approach in three public datasets composed of multimodal physiological data across different classification tasks: WESAD [47], the CSL-SHARE [48] and eSports Sensors (eSports) [49]. A brief dataset description is provided below, while a more comprehensive description can be found in Appendix B.

The WESAD dataset was introduced and made publicly available by Schmidt et al. [47]. Their study aimed to elicit different affective states in 15 participants. This multimodal dataset contains motion and physiological data collected by equipment placed on participants' wrists and chests. For our study, only chest sensors were used, which include a 3-axis accelerometer sensor (ACC) and physiological data from electrocardiogram (ECG), electrodermal activity (EDA), skin temperature (TEMP), electromyography (EMG) and respiration (RESP). The goal was to classify between three affective states, namely baseline, amused, and stressed conditions.

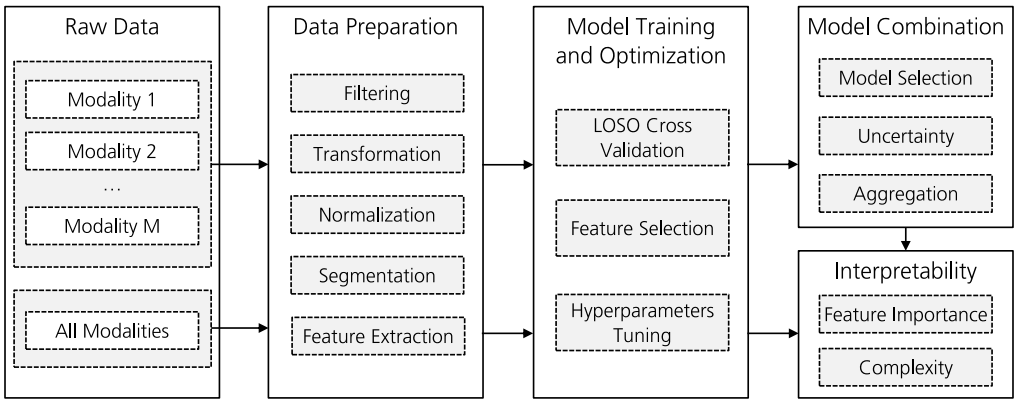


Fig. 2. A schematic representation of the machine learning pipeline, depicting the various stages of data preparation, model training and optimization, model combination, and model interpretability analysis.

The CSL-SHARE dataset contains activities of daily living and sports collected from 20 subjects and was made publicly available by Liu et al. [48]. Participants wore a knee bandage with two triaxial accelerometers (ACC), two triaxial gyroscope sensors (GYRO), four surface electromyography (sEMG) sensors, one biaxial electrogoniometer (GONIO), and one airborne microphone (MIC). The objective was to classify between 22 classes of daily living activities.

The eSports dataset was collected from 22 professional and amateur video game team matches during more than 40 h of recordings. While participants played video games, the following modalities were recorded: electromyography (EMG), electrodermal activity (EDA), heart rate (HRV), pulse oximetry (SPO2), and logs from the keyboard (KEYB) and mouse (MOUSE). The classification task was to distinguish between amateurs and professional players.

4.2. Experimental setup

Fig. 2 shows the pipeline used in this study. We trained each modality separately and developed a model using all modalities with early feature fusion. Data preprocessing varied for each modality and involved signal filtering to reduce noise and signal-specific transformation routines, e.g., extraction of RR intervals from ECG and the signal envelope from EMG. We also normalized some modalities to reduce the bias arising from inter-subject variability. Then, we extracted features for each window in the statistical, temporal, spectral, and fractal domains. Different modalities require dedicated feature sets due to the intrinsic properties of the data used in the experiments. Our experiments used physiological time series collected from different modalities (e.g., ECG, EMG, EDA). These modalities often require dedicated feature sets with features that allow for a more explicit representation of the important properties of the original signals and underlying physiological properties. A comprehensive description of the data preprocessing and feature extraction for each dataset is provided in Appendix C.1.

Regarding model training, we considered five models: Decision Tree, Random Forest, AdaBoost, Naive Bayes, and Support Vector Machines. For the WESAD and eSports, we used the Leave One Subject Out (LOSO) cross-validation. For the CSL-SHARE, we used a Group 5-fold cross-validation. A sequential feature selection and a grid search hyperparameter tuning were also applied for model optimization. To evaluate the performance of the classifiers, we used different measures depending on skewed class proportions. For WESAD, we used the $F1$-score with macro average, while for balanced datasets, particularly CSL-SHARE and eSports, we used the accuracy. The final model for each modality was the classifier with the best performance after the training and optimization. A comprehensive description of the selected models for each modality, their respective hyperparameters, and the number of selected features is provided in Appendix C.2.

Table 2
Performance evaluation of the best individual models and the best model learned with all the individual modalities (ALL). For the WESAD and eSports, the mean and standard deviation refers to LOSO; for the CSL-SHARE, they refer to Group 5-fold cross-validation. We also report the outcome of a random guesser based on the mathematical expectation of the results.

WESAD		CSL-SHARE		eSports	
Modality	$F1$-score	Modality	Accuracy	Modality	Accuracy
ACC	0.671 ± 0.178	ACC	0.685 ± 0.039	ACC	0.674 ± 0.154
EDA	0.616 ± 0.195	GYRO	0.728 ± 0.043	EDA	0.680 ± 0.245
TEMP	0.518 ± 0.188	GONIO	0.650 ± 0.035	HRV	0.602 ± 0.255
EMG	0.402 ± 0.126	EMG	0.375 ± 0.019	EMG	0.624 ± 0.232
ECG	0.555 ± 0.148	MIC	0.257 ± 0.015	SPO2	0.624 ± 0.162
RESP	0.613 ± 0.111			MOUSE	0.730 ± 0.147
				KEYB	0.749 ± 0.123
ALL	0.777 ± 0.155	ALL	0.835 ± 0.036	ALL	0.816 ± 0.110
Random	0.317	Random	0.045	Random	0.500

In the final step of the pipeline, the best models for each modality are combined using the late model fusion methods described in Section 2.2.1 weighted by the uncertainty measures described in Section 2.2.2. The feature importance for each model was measured using SHAP. We used the TreeExplainer [50] for the Decision Tree and Random Forest models and the KernelExplainer for the remaining models with $K = 100$ samples summarized using k-means. The feature importance and the complexity of explanations were measured for the individual models, the model learned with all modalities and the different model aggregation approaches.

4.3. Experimental results

The performance results of the best individual models and the best model learned with all the individual modalities for each dataset are presented in Table 2. We show the performance of a random guesser baseline in the last row based on the mathematical expectation of the results considering the number of classes and the class distribution of the observations.

A general observation for all the datasets is that the ALL model performs better than the individual modalities. It supports that the early fusion of individual modalities in these multimodal problems improves classification performance. The best individual modality for the WESAD was the ACC (0.671 ± 0.178), for the CSL-SHARE was the GYRO (0.728 ± 0.043), and for the eSports was the KEYB (0.749 ± 0.123). The lowest performer individual modality for the WESAD was the EMG (0.402 ± 0.126), for the CSL-SHARE was the MIC (0.257 ± 0.015), and for the eSports was the HRV (0.602 ± 0.255). Even though the individual modalities showed reduced performance, they all were above random guessing.

Table 3
Performance evaluation of baseline aggregation methods (unweighted) compared with their weighted version with different uncertainty measures. The best result per dataset and aggregation strategy is highlighted in bold. The performance refers to $F1$-score for the WESAD, whereas, for the CSL-SHARE and eSports refers to accuracy.

Aggregation	Weight	Performance		
		WESAD	CSL-SHARE	eSports
Majority voting	1	**0.752** ± 0.123	0.755 ± 0.035	0.777 ± 0.077
Sum rule	1	0.699 ± 0.137	**0.824** ± 0.032	**0.832** ± 0.061
Mean rule	1	0.699 ± 0.137	0.824 ± 0.032	0.832 ± 0.061
Max rule	1	0.625 ± 0.175	0.783 ± 0.034	0.793 ± 0.084
Prod rule	1	0.663 ± 0.211	0.806 ± 0.024	0.819 ± 0.075
Min rule	1	0.656 ± 0.209	0.706 ± 0.016	0.793 ± 0.084
Majority voting	AU_{max}	0.774 ± 0.130	**0.800** ± 0.030	0.840 ± 0.053
	AU_{entropy}	0.691 ± 0.146	0.792 ± 0.026	0.816 ± 0.075
	AU_{bayes}	0.697 ± 0.140	0.791 ± 0.025	0.816 ± 0.075
	EU_{vr}	**0.781** ± 0.131	0.786 ± 0.031	0.841 ± 0.054
	EU_{bayes}	0.772 ± 0.128	0.783 ± 0.023	0.833 ± 0.063
	TU[a]	0.776 ± 0.132	0.795 ± 0.030	**0.843** ± 0.053
	MR	0.765 ± 0.119	0.787 ± 0.026	0.819 ± 0.065
	MR_{rej}	0.766 ± 0.115	0.788 ± 0.026	0.833 ± 0.057
Sum rule	EU_{bayes}	0.705 ± 0.156	**0.827** ± 0.032	0.826 ± 0.067
	EU_{vr}	0.692 ± 0.155	0.823 ± 0.034	0.823 ± 0.082
	MR	**0.711** ± 0.153	0.827 ± 0.033	0.826 ± 0.066
	MR_{rej}	0.710 ± 0.150	0.826 ± 0.033	**0.827** ± 0.066

[a]Represents the combination of the best performing aleatoric and epistemic uncertainty measures.

In the following sections, we report the results of studying the late model fusion methods using UQ (Section 4.3.1), and the explanation complexity (Section 4.3.2).

4.3.1. Late model fusion methods

Table 3 reports the results of combining the individual models for each modality using the baseline aggregation methods and our proposed weighted version using uncertainty measures. Note that for soft aggregation strategies, only the best-performing aggregation method obtained during training was considered for the weighted variant of soft aggregation.

From Table 3, we can conclude that majority voting outperformed all equal-weighted soft voting strategies on the WESAD dataset. In contrasts, the opposite was observed for the CSL-SHARE and eSports datasets. As a result, the best weighted aggregation strategy for the WESAD dataset was majority voting weighted by EU_{vr}, while for the CSL-SHARE dataset, the sum rule weighted by EU_{bayes} achieved the highest score. In the case of the eSports dataset, although the unweighted version of the sum rule outperformed majority voting, the same did not hold true for the uncertainty-weighted versions. Here, majority voting weighted by $TU_{\{AU_{max};EU_{vr}\}}$ achieved higher performance than the weighted versions using the sum rule.

Overall, the weighted majority voting performed better than its unweighted counterpart across all datasets, except for AU_{entropy} and AU_{bayes} in the WESAD dataset. Concerning soft voting strategies, the use of uncertainty-based weighting led to similar performances compared to the unweighted approach.

Compared to models trained using all modalities, the top-performing weighted aggregation strategy achieved similar performance: (0.781 ± 0.131) for WESAD, (0.827 ± 0.032) for CSL-SHARE, and (0.843 ± 0.053) for eSports. In contrast, the models trained with all modalities obtained performance scores of (0.777 ± 0.155), (0.835 ± 0.036), and (0.816 ± 0.110), respectively. It is worth noting that aggregation methods with late fusion use less information than the models trained with an early fusion of all modalities, as they do not consider the dependence between modalities.

4.3.2. Explanation complexity

To study the explanation complexity, we focus on the top-performing aggregation strategy (both unweighted and weighted) for each dataset. Throughout the remaining analysis, we will refer to the best aggregation method for each dataset as BASE_{agg}. Specifically, BASE_{agg} corresponds to majority voting for the WESAD dataset, sum rule for CSL-SHARE, and eSports dataset. Furthermore, we will use UNC_{agg} to denote the best weighted aggregation method for each dataset. For instance, UNC_{agg} represents majority voting weighted by EU_{vr} for WESAD, sum rule weighted by EU_{bayes} for CSL-SHARE, and majority voting weighted by TU_{*} for eSports.

Fig. 3 shows the relationship between the mean explanation complexity and mean performance for the models learned on individual modalities, the models learned with all the available modalities (ALL), and the aggregation approaches. The details on the number of features and modalities are also present.

The models learned on individual modalities exhibited a similar relationship between performance and explanation complexity scores for the three datasets, demonstrating a consistent pattern. As expected, using individual modalities led to lower explanation complexities but with a decreased model performance. The mean explanation complexity for the ALL approach was similar in WESAD and CSL-SHARE datasets resulting in (0.78 ± 0.23) and (0.78 ± 0.18), respectively. For eSports, the complexity was lower (0.50 ± 0.34).

The ALL and BASE_{agg} share a similar behavior, with higher performance but increased explanation complexity. The BASE_{agg} attained slightly lower performance w.r.t. ALL in WESAD and CSL-SHARE and slightly higher performance in eSports. For the WESAD and eSports, the BASE_{agg} had a slightly lower explanation complexity than ALL. However, this pattern was not observed for CSL-SHARE.

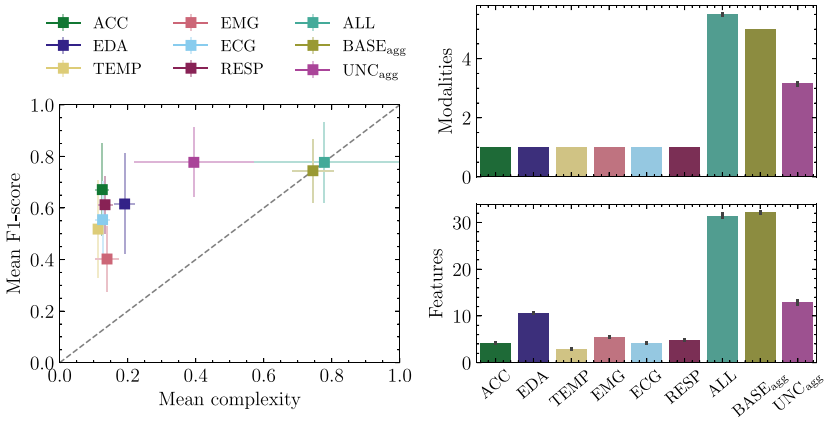
The UNC_{agg} model showed an interesting relationship, exhibiting similar performance to ALL but with a lower complexity score with an intermediate value between the individual models and ALL — (0.40 ± 0.18) for WESAD, (0.53 ± 0.21) for CSL-SHARE, and (0.29 ± 0.20) for eSports. This behavior was consistent among the three datasets. The UNC_{agg} achieved slightly higher performance than BASE_{agg} for WESAD and eSports, with a notably lower explanation complexity. For the CSL-SHARE, the mean accuracy was close to ALL. This lower complexity arises from UNC_{agg} using, on average, fewer modalities and features than ALL and BASE_{agg}, as illustrated in Fig. 3.

For the three datasets, the total number of modalities and features in the minimum relevant feature set used in the explanations was lower in UNC_{agg} w.r.t. ALL and BASE_{agg}.

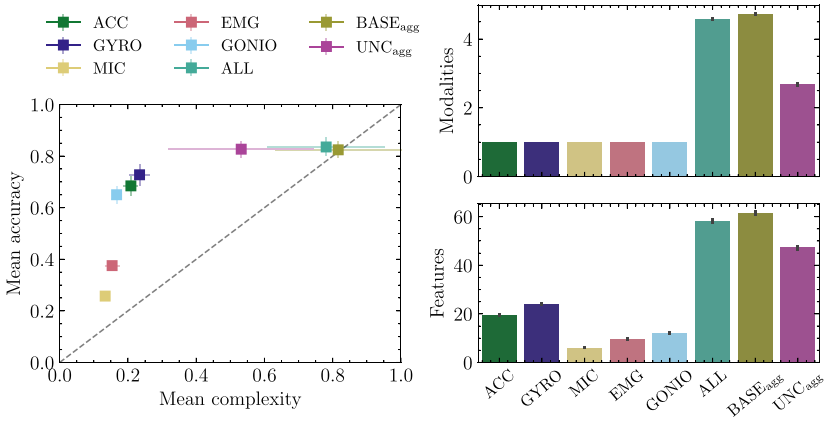
To calculate the number of modalities and features from the minimum relevant feature set, we use the threshold defined in Eq. (11). We studied the impact of this threshold on the results by analyzing how its value relates to the complexity. Specifically, we multiplied $\left|p(y|x_i) - \mathbb{E}[f(x_i)]\right|$ by the scalar α and calculated the resulting complexity for a range of values of α. The results are presented in Fig. 4. As expected, lower values of α lower the cardinality of the minimum relevant feature set, thus lowering the explanation complexity. Larger values of α lead to increased complexity. Generally, the growth rate for the number of features, modalities, and complexity stabilizes on a plateau for $\alpha \geq 1$. For the WESAD and eSports datasets, the number of features, modalities, and complexity was lower for UNC_{agg} compared to ALL, irrespective of α. Similar behavior was observed in the CSL-SHARE, except for the number of features, which remained approximately equal to ALL for $\alpha \leq 1$.

5. Discussion

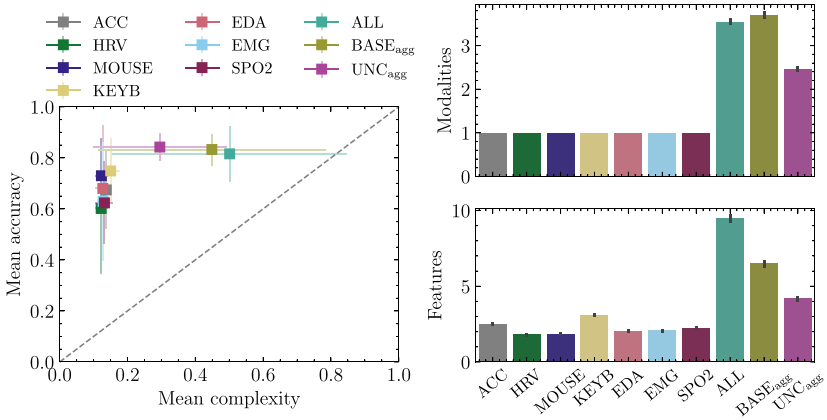
We propose a novel approach to lower the explanation complexity of feature-based time series models based on reducing the number of modalities and features used to explain multimodal data. Specifically, we use a late model fusion aggregation approach weighted by UQ



(a) WESAD



(b) CSL-SHARE



(c) eSports

Fig. 3. The left side plots show the relationship between the mean explanation complexity and mean performance for the individual modalities, ALL, and aggregation approaches. The horizontal and vertical error bars represent the standard deviation across all samples for explanation complexity and performance, respectively. The right side plots show the number of modalities and features of the minimum relevant feature subset. We report the median number of features and modalities and the error bar with a 95% confidence interval.

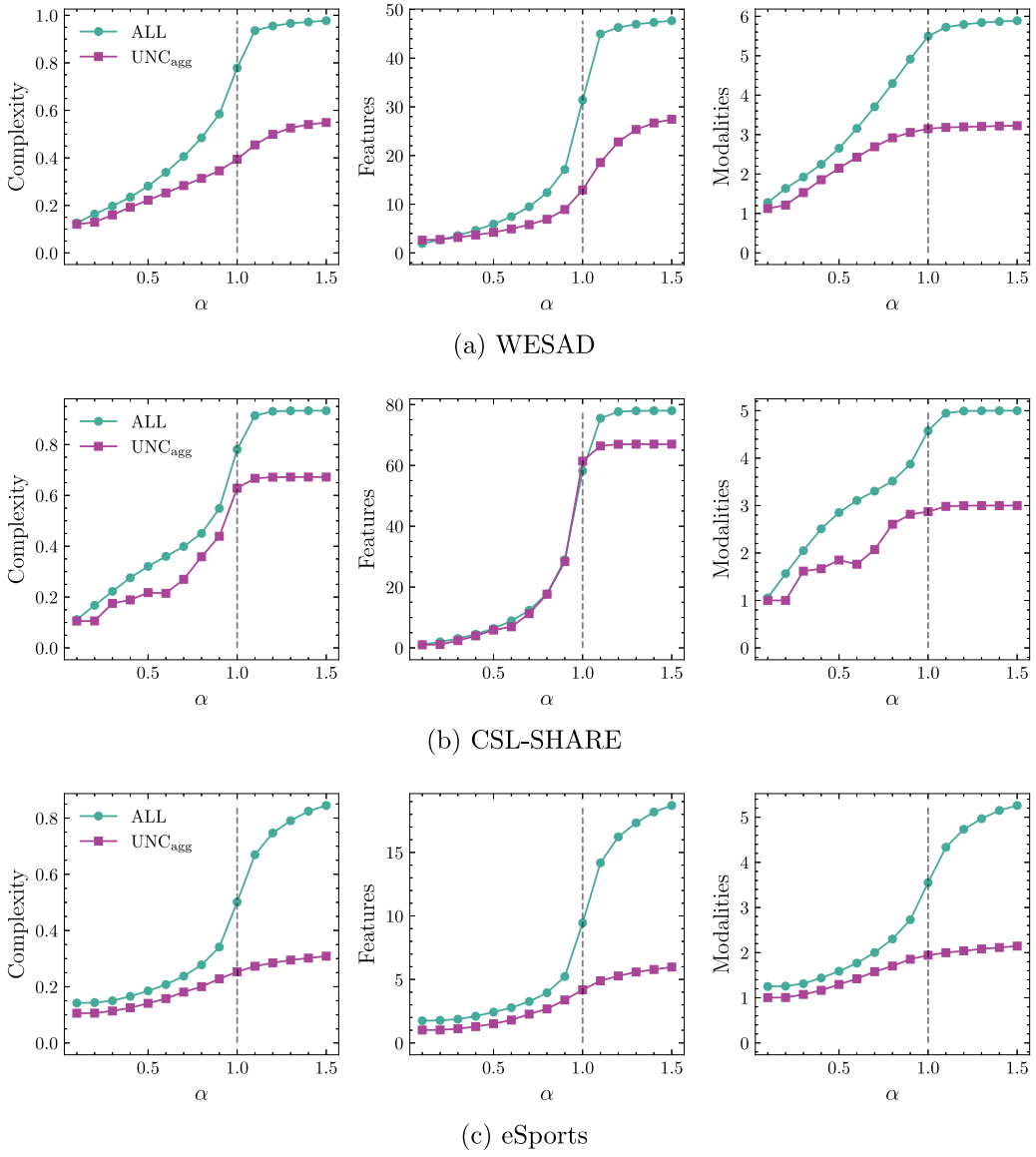


Fig. 4. The impact of the threshold in the minimum relevant feature size. We multiplied $|p(y|x) - \mathbb{E}[f(x)]|$ from Eq. (11) by the scalar α and calculated the resulting number of features, modalities, and complexity for a range of values of α. Each point represents the mean value across all samples. The dashed vertical line denotes the α used in our experiments. (a) WESAD dataset (b) CSL-SHARE dataset (c) eSports dataset.

measures that consider the most certain modalities to predict a sample. Consequently, this approach lowers the explanation complexity in terms of the number of features and modalities required to explain a given prediction. We argue that this reduction in explanation complexity yields less complex local explanations without compromising the models' performance.

Although not a main contribution of this work, it is important to emphasize that our proposed processing and modeling pipeline yielded improved results compared to Schmidt et al. [47], except for the ECG and RESP. For an extended discussion of this topic, refer to Appendix D.

A general assumption in previous literature is that linear models with fewer parameters or rule-based models with few rules are less complex than models with many parameters and rules [51,52]. Our work extends this assumption to the multimodal data setting, arguing that more concise explanations have fewer modalities and features. This notion is supported by previous research in social sciences [14,22], and user research studies focusing on model explainability [33]. In previous studies, metrics for post-hoc interpretability based on the number of features and interactions were proposed [11,31]. However, these methods are more generalized and do not account for the impact of modalities on the explanation. Therefore, our technical approach differs, where we introduce a novel measure for assessing model complexity appropriate for multimodal scenarios. The measure defined in Eq. (10) increases for explanations requiring more features and modalities.

We tested our approach with publicly available multimodal physiological datasets from different domains. The datasets differ in the number of modalities, the number of features, and also the number of classes (binary and multiclass settings). The UNC$_{agg}$ showed a consistent pattern of lower explanation complexity and approximately equal or slightly higher classifier performance w.r.t. the ALL baseline. The complexity reduction was higher in WESAD and CSL-SHARE than in eSports. The number of selected features during training and optimization for the eSports was reduced compared to the other two datasets. Thus, its complexity of explanation for the ALL started with a lower baseline value (0.41 ± 0.20) in comparison to the other two datasets (0.78 ± 0.23) and (0.78 ± 0.18). The potential of the proposed approach is maximized in scenarios where models require many features and modalities. This observation aligns with the fact that these models

will have higher baseline explanation complexity in the early fusion models.

We conducted an experiment to understand better the impact of the threshold that selects the minimum relevant feature set. In general, irrespective of its value, our proposed approach yields explanations with fewer modalities and features. The reduction in the number of features was more evident in the WESAD and eSports datasets and marginal in the CSL-SHARE. The complexity reduction on the CSL-SHARE was mostly attributed to the reduction in the number of modalities. Explaining an instance with the same number of features but referent to fewer modalities is simpler than with a high number of modalities.

Using an uncertainty-based late model fusion is a fundamental aspect of our approach, and it allows us to dynamically select the most reliable and certain models for each instance. It answers one of the challenges identified in the review of Zhang et al. [53] concerning how to properly aggregate multiple modalities for emotion recognition.

By incorporating uncertainty quantification measures as weights during the aggregation process, our method can adaptively assess the confidence level of each model's predictions. As a result, our approach outperforms static weighting strategies and yields better overall performance. The use of confidence or quality measures as weighted metrics for model combinations is supported by other studies in the literature [36,54].

While most weight measures used in majority voting aggregation outperformed the baseline majority voting with equal weights, the same cannot be said for soft voting aggregation. Although soft aggregation methods incorporate more information during the combination process, the predictions generated by different classifiers are only compatible if the output classifier scores represent well-calibrated probabilities. Otherwise, averaging or other combination methods may not produce meaningful results [39]. Future research will address the impact of individual model calibration on ensemble performance.

When comparing the performance improvements of aleatoric and epistemic uncertainty as weighted aggregation measures, we observed that epistemic uncertainty had a more pronounced effect on the WESAD dataset, while aleatoric uncertainty yielded higher scores on the CSL-SHARE. In the case of eSports, similar improvements were obtained using both aleatoric and epistemic uncertainty. Upon evaluating the amounts of epistemic and aleatoric uncertainty in the three datasets, we found that CSL-SHARE generally exhibited greater aleatoric uncertainty than WESAD and the opposite for epistemic uncertainty. For the eSports dataset, we observed that both aleatoric and epistemic uncertainty consistently had higher values compared to WESAD and CSL-SHARE. Therefore, our findings suggest that aleatoric and epistemic uncertainty measures improved the overall performance on datasets containing a higher proportion of aleatoric and epistemic uncertainty samples, respectively. While this conclusion may seem apparent, it emphasizes the importance of integrating both uncertainty sources for a more effective weighted aggregation strategy. Additionally, gaining a deeper understanding of the models during development can enhance the robustness and performance. We present the results of global aleatoric and epistemic uncertainty across all datasets in Appendix E.

Addressing missing data is an essential aspect of the practical utility of our proposed late fusion approach. In real-world applications, it is common for some modalities to be missing or to exhibit poor quality due to various factors, such as sensor failures or noisy environments. Our method offers a significant advantage in such scenarios, as it can still provide reliable predictions even when some modalities are missing, albeit with decreased performance. Our approach can dynamically adjust the weights assigned to the available modalities by leveraging the uncertainty-based weighting scheme. This capability increases our method's robustness and makes it more suitable for deployment in practical situations where data completeness cannot be guaranteed. In Appendix F we present an experimental analysis of model's performance with missing data.

While the current study provides valuable insights and advances, it is important to acknowledge its limitations. This work relies on calculating a representative feature set to reduce the complexity of the explanation. The method to select the representative feature set was based on a heuristic that measured the difference between the posterior and prior probabilities. Some features were ignored since their marginal contribution was deemed minor. As a consequence, the local accuracy of the explanation may have been reduced, but this approach provided a net benefit in terms of lower complexity. While the current study used a heuristic to select the threshold to determine the representative feature set, it is possible that a user research study could provide further insights into the needs of users and better inform the selection of the optimal feature set that might depend on several factors, such as the task at hand, user expertise, among others.

6. Conclusion

This work presents a novel approach to reduce the explanation complexity of feature-based time series models in multimodal data by lowering the number of modalities and features required to deliver local explanations. Our hypothesis is based on insights from social science studies that call for parsimony when delivering explanations.

The late model fusion weighted by uncertainty quantification measures creates a robust ensemble that favors the most certain modalities for each instance, decreasing explanation complexity without sacrificing model performance. Our late model fusion method outperforms static weighting strategies and yields better overall performance. The results suggested that model aggregation using uncertainty quantification helps reduce the complexity of local explanations, with the most significant reduction observed in datasets with numerous features and modalities.

Uncertainty quantification and explainable AI are crucial in ensuring the reliability and transparency of AI systems. Our work serves as a compelling example of how these research topics can be effectively integrated to enhance overall performance, leading to more reliable and successful deployments.

CRediT authorship contribution statement

Duarte Folgado: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Funding acquisition. **Marília Barandas:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Funding acquisition. **Lorenzo Famiglini:** Formal analysis, Writing – review & editing. **Ricardo Santos:** Formal analysis, Writing – review & editing. **Federico Cabitza:** Writing – review & editing, Supervision. **Hugo Gamboa:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have used publicly available datasets.

Acknowledgments

This work was supported by European funds through the Recovery and Resilience Plan, project "Center for Responsible AI", project number C645008882-00000055.

Appendix A. A practical note on measuring explanation complexity from feature importance

Let us consider a sample taken from the illustrative dataset of Fig. 1, composed of 7 features from 3 modalities:

$$R = \{d_{emg1},\ d_{emg2},\ d_{eda1},\ d_{eda2},\ d_{eda3},\ d_{acc1},\ d_{acc2}\}\ ,|R| = 7\ .$$

$$M = \{EMG,\ EDA,\ ACC\}\ ,|M| = 3\ .$$

The minimum relevant feature set (calculated using Eq. (11)) required to explain the sample is as follows:

$$r = \{d_{emg1},\ d_{emg2},\ d_{acc1},\ d_{acc2}\}\ ,|r| = 4\ .$$

$$m = \{EMG,\ ACC\}\ ,|m| = 2\ .$$

Therefore, the complexity of the local explanation of this sample measured according to Eq. (10) is given as:

$$c = \frac{1}{2}\left(\frac{|r|}{|R|} + \frac{|m|}{|M|}\right) = \frac{1}{2}\left(\frac{4}{7} + \frac{2}{3}\right)\ .$$

Appendix B. Dataset description

The following paragraphs describe the three datasets used in our experiments. Although the description contains a complete dataset presentation, we selected some modalities or sensor positions in some cases. The modalities and positions considered in our experiments are the ones described in the sensing modalities. This procedure aims to simplify our experiments and results presentation and does not changes the validity of our conclusions on feature-based explanations.

B.1. Wearable Stress and Affect Detection (WESAD)

Description: WESAD is a publicly available wearable stress and affect detection dataset. This multimodal dataset contains physiological and motion data recorded from both a wrist- and a chest-worn devices. The dataset contains three different affective states (neutral, stress, and amusement). In addition, self-reports of the subjects, which were obtained using several established questionnaires, are contained in the dataset [47].

Task: The goal was to distinguish between the neutral, amusement, and stress emotion states. Therefore, it was a three-class classification problem.

Sample: Twelve male subjects and 3 female subjects. The mean age was 27.5 ± 2.4 years (N = 15).

Sensing modalities:

- Electromyography — BiosignalsPlux (Plux, Lisbon, Portugal) at 700 Hz.
- Eletrodermal activity — RespiBAN Professional (Plux, Lisbon, Portugal) at 700 Hz.
- Temperature — RespiBAN Professional (Plux, Lisbon, Portugal) at 700 Hz.
- Inertial Measurement Unit — BiosignalsPlux (Plux, Lisbon, Portugal) at 700 Hz.
- Respiration — RespiBAN Professional (Plux, Lisbon, Portugal) at 700 Hz.
- Electrocardiogram — BiosignalsPlux (Plux, Lisbon, Portugal) at 700 Hz.

The RespiBAN was placed around the subject's chest. The respiration was recorded using an respiration inductive plethysmography. The ECG data was recorded using a standard three-point ECG. In order to allow the subject to move as freely as possible, the EDA signal was recorded on the *rectus abdominis*, and the TEMP sensor was placed on the sternum. The EMG data were recorded on the *upper trapezius* muscle on both sides of the spine.

B.2. Cognitive Systems Lab Sensor-based Human Activity REcordings (CSL-SHARE)

Description: The CSL-SHARE dataset covers 22 types of activities of daily living and sports in a total of 691 min, of which 363 min are segmented and annotated. In this dataset, the authors used two triaxial accelerometers, two triaxial gyroscopes, four surface electromyography (sEMG) sensors, one biaxial electrogoniometer, and one airborne microphone integrated into a knee bandage [48].

Task: The goal was to distinguish between several human activity tasks. There were 22 classes.

Sample: Five females and 15 males. The mean age was 30.5 ± 5.8 (N = 20).

Sensing modalities:

All the following modalities were acquired at 1000 Hz using the BiosignalsPlux (Plux, Lisbon, Portugal): accelerometer, gyroscope, eletromiography, eletrogoniometer, and Airborne microphone.

The accelerometers and gyroscopes were placed on a knee bandage in the thigh and shank. The electromyography sensors were placed at *musculus vastus medialis*, *musculus tibialis anterior*, *musculus biceps femoris*, and *musculus gastrocnemius*. The goniometer and airborne microphone were on the right leg.

B.3. eSports Sensors (eSports)

Description: Data collected during an experiment with professional and amateur teams in 22 video game matches with more than 40 h of recordings. Recorded data include the players' physiological activity, e.g. movements, pulse, saccades, self-reported aftermatch surveys, and in-game data [49].

Task: The goal was to distinguish the player's skill between professional and amateur. Therefore, it was a binary classification problem.

Sample: Five amateur and 5 professional players (N = 10).

Sensing modalities:

- Electromyography — Grove EMG Sensor v1.11 (Seed Studio, Shenzhen, China) at 36 Hz.
- Eletrodermal activity — Grove GSR Sensor v1.22 (Seed Studio, Shenzhen, China) at 36 Hz.
- Inertial Sensor Unit — Bosch BNO055 (Bosch, Gerlingen, Germany) at 36 Hz. The sensors was located on the wrists on both hands. We only used the linear acceleration.
- Input (keyboard and mouse) logger data running on the participant's PC. These data are indirect indicators of the hand movement activity.
- Heart rate — Polar OH18 armband (Polar Electro Oy, Kempele, Finland) at 1 Hz.

These data were processed to remove the noise and outliers. Outliers for each sensor were removed by clipping values to an interval between 0.5 and 99.5 percentiles. Then the signal was smoothed using an exponential moving average with a half-live value of 1 s. Finally, all signals were resampled to 1 Hz by averaging or summation, depending on the nature of the source sensor data.

Appendix C. Experimental setup

In the context of time series classification, the features that have the most predictive performance depend on the specific data and questions being asked. Adding domain-specific features, in addition to or instead of general-purpose feature sets, can improve the interpretability of the results. This section describes the feature sets used for each dataset. We also provide the models, hyperparameters, and the total number of selected features for each experiment to promote reproducibility.

Table C.4
Selected models for each modality with corresponding hyperparameters and the number of selected features for WESAD dataset. The hyperparameters not referenced were set to the default values of scikit-learn implementation.

Modality	Classifier	Hyperparameters	# Features
ACC	Random Forest	*n_estimators = 100* *max_depth = 4* *min_samples_split = 20* *class_weight = 'balanced'*	7
EDA	Naive Bayes	–	14
TEMP	Random Forest	*n_estimators = 100* *max_depth = 3* *min_samples_split = 20* *class_weight = 'balanced'*	4
EMG	SVM	*kernel = 'sigmoid'* *gamma = 0.1* *C = 0.1* *class_weight = 'balanced'*	10
ECG	Random Forest	*n_estimators = 100* *max_depth = 5* *min_samples_split = 20* *class_weight = 'balanced'*	7
RESP	Random Forest	*n_estimators = 100* *max_depth = 3* *min_samples_split = 20* *class_weight = 'balanced'*	7
ALL	SVM	*kernel = 'sigmoid'* *gamma = 0.001* *C = 0.1* *class_weight = 'balanced'*	49

Table C.5
Selected models for each modality with corresponding hyperparameters and the number of selected features for CSL-SHARE dataset. The hyperparameters not referenced were set to the default values of scikit-learn implementation.

Modality	Classifier	Hyperparameters	# Features
ACC	Random Forest	*n_estimators = 200* *max_depth = 9* *min_samples_split = 30*	23
GYRO	Random Forest	*n_estimators = 200* *max_depth = 10* *min_samples_split = 15*	30
GONIO	Random Forest	*n_estimators = 200* *max_depth = 10* *min_samples_split = 15*	14
EMG	Random Forest	*n_estimators = 200* *max_depth = 9* *min_samples_split = 15*	15
MIC	Random Forest	*n_estimators = 200* *max_depth = 6* *min_samples_split = 40*	8
ALL	Random Forest	*n_estimators = 200* *max_depth = 10* *min_samples_split = 15*	78

C.1. Preprocessing and feature extraction

C.1.1. WESAD dataset

The signal preprocessing applied to each signal modality of the WESAD dataset was as follows:

- **ACC**: Magnitude of 3-axial accelerometer.
- **EDA**: The signal was filtered using a 2nd order lowpass Butterworth filter with cutoff of 5 Hz. Afterward, the decomposition into phasic and tonic components was performed using *cvxEDA* [55]. Then, a Min-Max normalization was applied to reduce the bias resulting from inter-subject variability.
- **TEMP**: A moving average filter with window size of $2 \times f_s$ was applied to reduce signal noise. A Min-Max normalization was applied to reduce the bias resulting from inter-subject variability.
- **EMG**: A baseline removal and a 3^{th} order 10–350 Hz bandpass Butterworth were applied.
- **ECG**: Interbeat interval was detected using the neuroKit2 algorithm for peak detection [56].
- **RESP**: The signal was filtered using a 2nd order 0.1–0.35 Hz bandpass Butterworth filter and a constant detrending afterward.

For feature extraction, TSFEL [57] was used to extract statistical and temporal features of the ACC, EDA, and TEMP modalities. For EMG, we followed the study of Phinyomark et al. [58] to extract statistical, temporal, and spectral domain features of the EMG signal. For the ECG, time, frequency, and non-linear domain features of the interbeat interval were extracted using NeuroKit2 [56]. Finally, for the RESP modality, statistical domain features of inspiration and expiration cycles were extracted.

C.1.2. CSL-SHARE dataset

In the CSL-SHARE dataset, the magnitude of ACC and GYRO modalities was computed. For the other three modalities (GONIO, EMG, and

Table C.6
Selected models for each modality with corresponding hyperparameters and the number of selected features for eSports dataset. The hyperparameters not referenced were set to the default values of scikit-learn implementation.

Modality	Classifier	Hyperparameters	# Features
ACC	Random Forest	*n_estimators = 102* *max_depth = 4* *min_samples_split = 31* *min_samples_leaf = 12*	3
EDA	Decision Tree	*max_depth = 4* *min_samples_split = 15* *max_leaf_nodes = 11* *min_samples_leaf = 17*	3
HRV	Decision Tree	*max_depth = 3* *min_samples_split = 12* *max_leaf_nodes = 37* *min_samples_leaf = 12*	3
EMG	Random Forest	*n_estimators = 237* *max_depth = 3* *min_samples_split = 37* *min_samples_leaf = 19*	3
SPO2	Decision Tree	*max_depth = 4* *min_samples_split = 30* *max_leaf_nodes = 38* *min_samples_leaf = 17*	4
MOUSE	Random Forest	*n_estimators = 102* *max_depth = 4* *min_samples_split = 31* *min_samples_leaf = 12*	3
KEYB	Random Forest	*n_estimators = 100* *max_depth = 4* *min_samples_split = 12* *min_samples_leaf = 12*	4
ALL	SVM	*kernel = 'sigmoid'* *gamma = 0.1* *C = 0.001*	22

MIC), a baseline removal was applied. For EMG, a 3rd-order 10–350 Hz bandpass Butterworth filter was also applied.

For feature extraction, Catch22 [59], which extracts 22 features from the statistical, temporal, and spectral domains, was used for ACC, GYRO, GONIO, and MIC modalities. For EMG, we followed the study of Phinyomark et al. [58] to extract statistical, temporal, and spectral domain features of the EMG signal.

C.1.3. Esports dataset

In the eSports dataset, the magnitude of linear acceleration was computed for the ACC modality. No further preprocessing was applied to the other modalities, as the dataset was already preprocessed, as mentioned in Appendix B.3. For feature extraction, TSFEL [57] was used to extract statistical and temporal features from all modalities.

C.2. Model training and optimization

See Tables C.4–C.6.

Appendix D. Additional discussion on WESAD results

Section 4 provided the experimental evaluation of our proposed classifiers. Previous works from the literature only reported the results for the model learned with all modalities. However, for the WESAD, the authors provided results per individual modality. Therefore, this section provides a further discussion of our results compared with the performance reported by Schmidt et al. [47] (see Table D.7).

Our proposed approach attained higher performance for ACC, EDA, TEMP, and EMG. The results from [47] were slightly higher in the ECG and RESP. The early fusion model with all the modalities was better in our approach (0.777 ± 0.155). The higher $F1$-score can be explained by changes we adopted to the original pipeline design.

Table D.7
Performance evaluation of the best models for each modality comparing the results from Schmidt et al. [47] and ours. $F1$-score with macro average is used as the performance measure. We report the mean and standard deviation of LOSO. The standard deviation of the LOSO was not reported by Schmidt et al. [47]. The best results are highlighted in bold.

	WESAD	
	Schmidt et al. [47]	Ours
ACC	0.443	**0.671** ± 0.178
EDA	0.483	**0.616** ± 0.195
TEMP	0.425	**0.518** ± 0.188
EMG	0.381	**0.402** ± 0.126
ECG	**0.560**	0.555 ± 0.148
RESP	**0.618**	0.613 ± 0.111
ALL	0.725	**0.777** ± 0.155

We used a larger window shift of 6 s, as we observed no significant variations in the $F1$-score following multiple tests, reducing the training time required for the model. During preprocessing, we normalized feature values for some modalities per subject. This approach aided in reducing model overfitting and eliminating residual bias stemming from differences in basal subject-specific values. It is important to note differences in the feature set used in our study and that of Schmidt et al. [47]. Specifically, we utilized open-source libraries to extract ECG, RESP, and EDA data [56,57]. Our LOSO's standard deviation was moderate, which reflects some inter-subject variability. Nevertheless, this value was consistent with findings from other studies utilizing the WESAD dataset [60].

Appendix E. Aleatoric and epistemic uncertainty for model rejection

When comparing the performance improvements of aleatoric and epistemic uncertainty as weighted aggregation measures, we observed

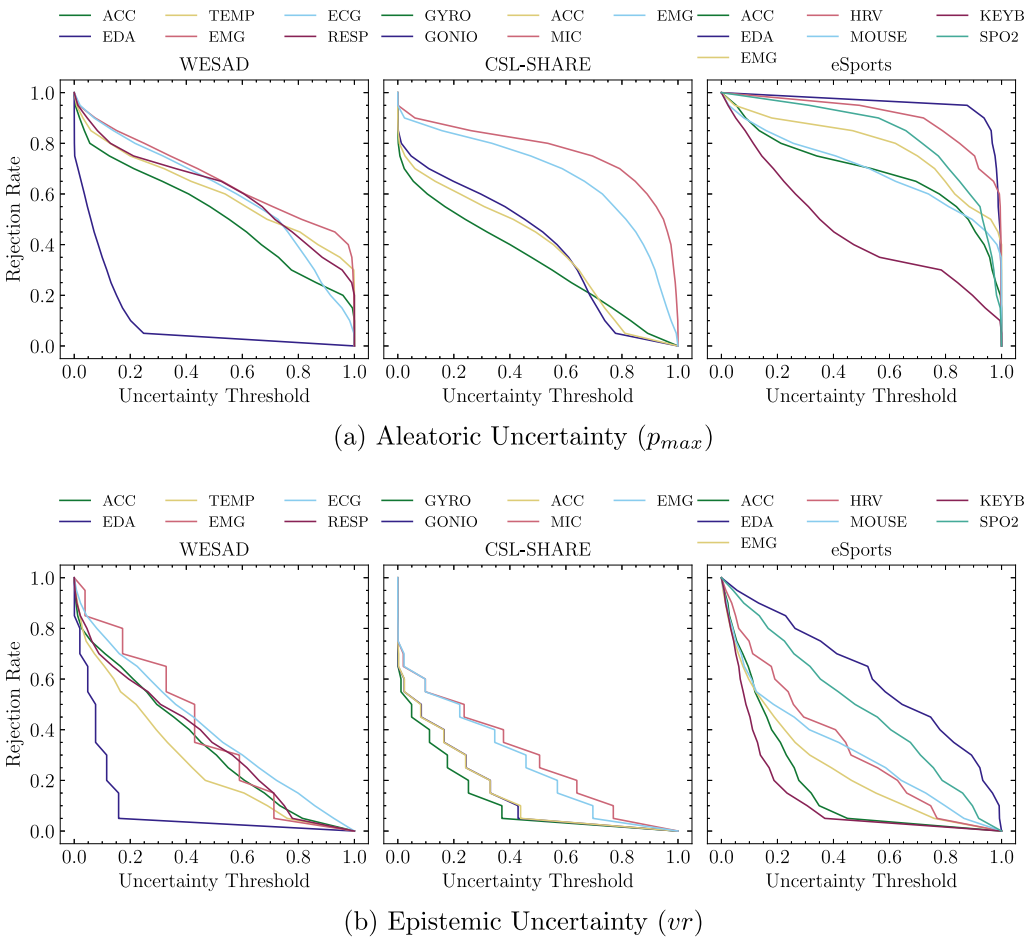


Fig. E.5. Rejection rate as a function of normalized uncertainty values for (a) aleatoric and (b) epistemic uncertainties. The uncertainty threshold is normalized to the maximum theoretical value of each uncertainty measure. Results are shown for the three datasets using the best-performing uncertainty measures when applying majority voting, with p_{max} for aleatoric uncertainty and vr for epistemic uncertainty.

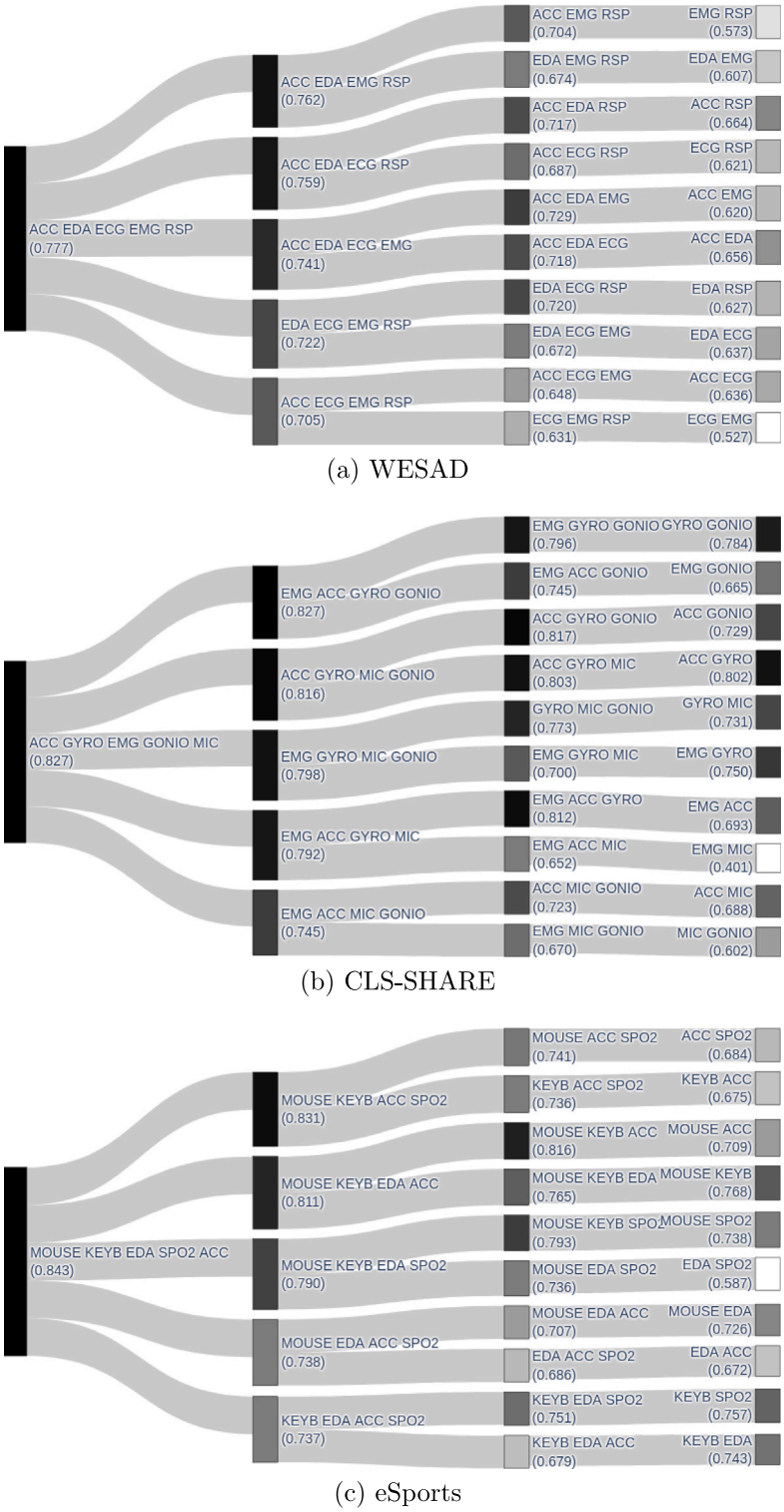
that epistemic uncertainty had a more pronounced effect on the WESAD dataset, while aleatoric uncertainty led to higher scores on the CSL-SHARE dataset. In the case of eSports, similar improvements were obtained using both aleatoric and epistemic uncertainty. In fact, the best-performing aggregation was achieved using a combination of both aleatoric and epistemic uncertainty.

Fig. E.5 displays the rejection rate as a function of the dataset uncertainty values, which are normalized by their maximum theoretical value. For instance, when the uncertainty threshold is set to 1, it corresponds to the maximum theoretical value, and as a result, no samples are rejected, leading to a rejection rate of zero. The figure presents the results for all datasets using the best-performing uncertainty measures when applying majority voting, namely p_{max} for aleatoric uncertainty and vr for epistemic uncertainty.

Upon evaluating the amounts of epistemic and aleatoric uncertainty in the three datasets, we found that CSL-SHARE generally exhibited greater aleatoric uncertainty than WESAD and the opposite for epistemic uncertainty. In fact, CSL-SHARE had two modalities (MIC and EMG) with high aleatoric uncertainty throughout the entire dataset. This observation may have contributed to the differences in the obtained results of the two datasets. For the eSports dataset, we observed that both aleatoric and epistemic uncertainty consistently had higher values compared to WESAD and CSL-SHARE. This fact may have contributed to both uncertainty measures obtaining similar improvements.

Appendix F. Missing data

In Fig. F.6, we present an analysis of the model's performance with missing data, focusing on the best-performing model from Table 3, i.e., the majority voting weighted by EU_{vr} for WESAD dataset, the sum rule weighted by EU_{bayes} for CSL-SHARE dataset and the majority voting weighted by $TU_{\{AU_{max};EU_{vr}\}}$ for eSports dataset. The diagram depicts the model's performance with varying combinations of modalities, using grayscale to represent the obtained performance (dark for highest and white for lowest). For simplicity, the analysis begins with five modalities (instead of the available six for WESAD and seven for eSports) and concludes with the combination of two modalities. For the WESAD and eSports datasets, the modalities with the lowest performance are not included in the diagram, i.e., the TEMP in WESAD and the HRV and EMG in eSports. The drop in performance varies depending on the missing modality. For example, in WESAD dataset, the removal of the ECG modality results in a minor performance decrease from 0.777 to 0.762, whereas the removal of the EDA modality leads to a more significant drop from 0.777 to 0.705. While the combination of various models should theoretically enhance the performance of individual models, there are cases where the opposite effect was observed. For instance, in the eSports dataset, combining EDA and SPO2 models resulted in a performance score of 0.587, lower than their individual scores of 0.680 and 0.624, respectively. Similarly, in the CSL-SHARE dataset, the fusion of MIC and GONIO models led to a performance of 0.602. In this case, GONIO's individual performance was higher at 0.650, while MIC's performance was lower at 0.257.



(a) WESAD

(b) CLS-SHARE

(c) eSports

Fig. F.6. Diagram illustrating the performance of the best-performing model with uncertainty-based aggregation methods when handling missing modalities. The gray scale indicates the obtained performance for each combination of modalities (black for highest, white for lowest). The diagram starts with five modalities and ends with a combination of two modalities. The performance values for each combination are presented alongside the modalities in the figure.

References

[1] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[2] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Inf. Fusion 77 (2022) 29–52, http://dx.doi.org/10.1016/j.inffus.2021.07.016.

[3] C. Dindorf, W. Teufl, B. Taetz, G. Bleser, M. Fröhlich, Interpretability of input representations for gait classification in patients after total hip arthroplasty, Sensors 20 (16) (2020) 4385, http://dx.doi.org/10.3390/s20164385.

[4] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ECGs, Comput. Biol. Med. 133 (2021) 104393, http://dx.doi.org/10.1016/j.compbiomed.2021.104393.

[5] C. Dindorf, J. Konradi, C. Wolf, B. Taetz, G. Bleser, J. Huthwelker, F. Werthmann, E. Bartaguiz, J. Kniepert, P. Drees, U. Betz, M. Fröhlich, Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (XAI), Sensors 21 (18) (2021) 6323, http://dx.doi.org/10.3390/s21186323.

[6] N. Mollaei, C. Fujão, L. Silva, J. Rodrigues, C. Cepeda, H. Gamboa, Human-centered explainable artificial intelligence: Automotive occupational health protection profiles in prevention musculoskeletal symptoms, Int. J. Environ. Res. Public Health 19 (15) (2022) 9552, http://dx.doi.org/10.3390/ijerph19159552.

[7] N. Rahim, S. El-Sappagh, S. Ali, K. Muhammad, J. Del Ser, T. Abuhmed, Prediction of Alzheimer's progression based on multimodal Deep-learning-based fusion and visual Explainability of time-series data, Inf. Fusion 92 (2023) 363–388, http://dx.doi.org/10.1016/j.inffus.2022.11.028.

[8] M.Z. Uddin, A. Soylu, Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning, Sci. Rep. 11 (1) (2021) 16455, http://dx.doi.org/10.1038/s41598-021-95947-y.

[9] N. Bussmann, P. Giudici, D. Marinelli, J. Papenbrock, Explainable machine learning in credit risk management, Comput. Econ. 57 (1) (2021) 203–216, http://dx.doi.org/10.1007/s10614-020-10042-0.

[10] F. Oviedo, J.L. Ferres, T. Buonassisi, K.T. Butler, Interpretable and explainable machine learning for materials science and chemistry, Acc. Mater. Res. 3 (6) (2022) 597–607, http://dx.doi.org/10.1021/accountsmr.1c00244.

[11] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C.A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, General pitfalls of model-agnostic interpretation methods for machine learning models, in: xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, Springer, 2022, pp. 39–68, http://dx.doi.org/10.1007/978-3-031-04083-2_4.

[12] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239, http://dx.doi.org/10.1109/34.667881.

[13] M. Mohandes, M. Deriche, S.O. Aliyu, Classifiers combination techniques: A comprehensive review, IEEE Access 6 (2018) 19626–19639, http://dx.doi.org/10.1109/ACCESS.2018.2813079.

[14] T. Lombrozo, Explanatory preferences shape learning and inference, Trends in Cognitive Sciences 20 (10) (2016) 748–759, http://dx.doi.org/10.1016/j.tics.2016.08.001.

[15] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, 2017, pp. 618–626, http://dx.doi.org/10.1109/ICCV.2017.74.

[16] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognit. Lett. 150 (2021) 228–234, http://dx.doi.org/10.1016/j.patrec.2021.06.030.

[17] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[18] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 1–10.

[19] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv:1702.08608 [cs, stat], URL: http://arxiv.org/abs/1702.08608, arXiv:1702.08608.

[20] J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (5) (2021) 593, http://dx.doi.org/10.3390/electronics10050593.

[21] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, ACM Comput. Surv. (2023) http://dx.doi.org/10.1145/3583558.

[22] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inform. 113 (2021) 103655, http://dx.doi.org/10.1016/j.jbi.2020.103655.

[23] I. Askira-Gelman, Knowledge discovery: comprehensibility of the results, in: Proceedings of the Thirty-First Hawaii International Conference on System Sciences, Vol. 5, IEEE, 1998, pp. 247–255, http://dx.doi.org/10.1109/HICSS.1998.648319.

[24] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 253–263, URL: https://aclanthology.org/I17-1026.

[25] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, Mach. Learn. 102 (2016) 349–391, http://dx.doi.org/10.1007/s10994-015-5528-6.

[26] K.P. Burnham, D.R. Anderson, Multimodel inference: understanding AIC and BIC in model selection, Sociol. Methods Res. 33 (2) (2004) 261–304, http://dx.doi.org/10.1177/0049124104268644.

[27] L. Zhao, Q. Hu, W. Wang, Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso, IEEE Trans. Multimed. 17 (11) (2015) 1936–1948, http://dx.doi.org/10.1109/TMM.2015.2477058.

[28] G. Plumb, M. Al-Shedivat, Á.A. Cabrera, A. Perer, E. Xing, A. Talwalkar, Regularizing black-box models for improved interpretability, Adv. Neural Inf. Process. Syst. 33 (2020) 10526–10536.

[29] S.M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, G. Parker, Interpretation of depression detection models via feature selection methods, IEEE Trans. Affect. Comput. (2020) http://dx.doi.org/10.1109/taffc.2020.3035535.

[30] S.L. Buchner, Multimodal Feature Selection to Unobtrusively Model Trust, Workload, and Situation Awareness (Ph.D. thesis), University of Colorado at Boulder, 2022.

[31] U. Bhatt, A. Weller, J.M.F. Moura, Evaluating and aggregating feature-based model explanations, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 2020, pp. 3016–3022, http://dx.doi.org/10.24963/ijcai.2020/417.

[32] R.W. Batterman, C.C. Rice, Minimal model explanations, Philos. Sci. 81 (3) (2014) 349–376, http://dx.doi.org/10.1086/676677.

[33] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S.J. Gershman, F. Doshi-Velez, Human evaluation of models built for interpretability, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7, 2019, pp. 59–67, http://dx.doi.org/10.1609/hcomp.v7i1.5280.

[34] E. Fersini, F.A. Pozzi, E. Messina, Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers, in: 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA, IEEE, 2015, pp. 1–8.

[35] M. Shahhosseini, G. Hu, H. Pham, Optimizing ensemble weights and hyperparameters of machine learning models for regression problems, Mach. Learn. Appl. 7 (2022) 100251, http://dx.doi.org/10.1016/j.mlwa.2022.100251, URL: https://www.sciencedirect.com/science/article/pii/S2666827022000020.

[36] N. Poh, J. Kittler, A unified framework for biometric expert fusion incorporating quality measures, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2011) 3–18, http://dx.doi.org/10.1109/TPAMI.2011.102.

[37] M. Barandas, D. Folgado, R. Santos, R. Simão, H. Gamboa, Uncertainty-based rejection in machine learning: Implications for model development and interpretability, Electronics 11 (3) (2022) 396, http://dx.doi.org/10.3390/electronics11030396.

[38] S. Chitroub, Classifier combination and score level fusion: concepts and practical aspects, Int. J. Image Data Fusion 1 (2) (2010) 113–135, http://dx.doi.org/10.1080/19479830903561944.

[39] A. Tornede, L. Gehring, T. Tornede, M. Wever, E. Hüllermeier, Algorithm selection on a meta level, Mach. Learn. (2022) 1–34, http://dx.doi.org/10.1007/s10994-022-06161-4.

[40] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Mach. Learn. 110 (2021) 457–506, http://dx.doi.org/10.1007/s10994-021-05946-3.

[41] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, S. Udluft, Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 1184–1193.

[42] J. Mena, O. Pujol, J. Vitrià, Uncertainty-based rejection wrappers for black-box classifiers, IEEE Access 8 (2020) 101721–101746, http://dx.doi.org/10.1109/ACCESS.2020.2996495.

[43] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67.

[44] H. Chen, I.C. Covert, S.M. Lundberg, S.-I. Lee, Algorithms to estimate Shapley value feature attributions, Nat. Mach. Intell. (2023) 1–12.

[45] G. Petelin, G. Cenikj, T. Eftimov, Towards understanding the importance of time-series features in automated algorithm performance prediction, Expert Syst. Appl. 213 (2023) 119023.

[46] J. Bento, P. Saleiro, A.F. Cruz, M.A. Figueiredo, P. Bizarro, Timeshap: Explaining recurrent models through sequence perturbations, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 2565–2573.

[47] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing WESAD, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 400–408, http://dx.doi.org/10.1145/3242969.3242985.

[48] H. Liu, Y. Hartmann, T. Schultz, CSL-SHARE: A multimodal wearable sensor-based human activity dataset, Front. Comput. Sci. 3 (2021) 759136, http://dx.doi.org/10.3389/fcomp.2021.759136.

[49] A. Smerdov, B. Zhou, P. Lukowicz, A. Somov, Collection and validation of psychophysiological data from professional and amateur players: a multimodal esports dataset, 2020, arXiv preprint arXiv:2011.00958.

[50] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 2522–5839, http://dx.doi.org/10.1038/s42256-019-0138-9.

[51] A.A. Freitas, Comprehensible classification models: a position paper, ACM SIGKDD Explor. Newsl. 15 (1) (2014) 1–10, http://dx.doi.org/10.1145/2594473.2594475.

[52] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decis. Support Syst. 51 (1) (2011) 141–154, http://dx.doi.org/10.1016/j.dss.2010.12.003.

[53] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, Inf. Fusion 59 (2020) 103–126, http://dx.doi.org/10.1016/j.inffus.2020.01.011.

[54] A. Campagner, D. Ciucci, F. Cabitza, Aggregation models in ensemble learning: A large-scale comparison, Inf. Fusion 90 (2023) 241–252, http://dx.doi.org/10.1016/j.inffus.2022.09.015.

[55] A. Greco, G. Valenza, A. Lanata, E.P. Scilingo, L. Citi, cvxEDA: A convex optimization approach to electrodermal activity processing, IEEE Trans. Biomed. Eng. 63 (4) (2015) 797–804, http://dx.doi.org/10.1109/TBME.2015.2474131.

[56] D. Makowski, T. Pham, Z.J. Lau, J.C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S.A. Chen, NeuroKit2: A python toolbox for neurophysiological signal processing, Behav. Res. Methods (2021) 1–8, http://dx.doi.org/10.3758/s13428-020-01516-y.

[57] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, H. Gamboa, TSFEL: Time series feature extraction library, SoftwareX 11 (2020) 100456, http://dx.doi.org/10.1016/j.softx.2020.100456.

[58] A. Phinyomark, P. Phukpattaranont, C. Limsakul, Feature reduction and selection for EMG signal classification, Expert Syst. Appl. 39 (8) (2012) 7420–7431, http://dx.doi.org/10.1016/j.eswa.2012.01.102.

[59] C.H. Lubba, S.S. Sethi, P. Knaute, S.R. Schultz, B.D. Fulcher, N.S. Jones, catch22: CAnonical Time-series CHaracteristics: Selected through highly comparative time-series analysis, Data Min. Knowl. Discov. 33 (6) (2019) 1821–1852, http://dx.doi.org/10.1007/s10618-019-00647-x.

[60] M. Yan, Z. Deng, B. He, C. Zou, J. Wu, Z. Zhu, Emotion classification with multichannel physiological signals using hybrid feature and adaptive decision fusion, Biomed. Signal Process. Control 71 (2022) 103235, http://dx.doi.org/10.1016/j.bspc.2021.103235.