

Interpreting Stress Detection Models using SHAP and Attention for MuSe-Stress 2022

Ho-min Park, Ganghyun Kim, Jinsung Oh, Arnout Van Messem, and Wesley De Neve

Abstract—Understanding emotional reactions, especially stress, during job interviews holds significant implications for assessing the well-being of candidates and tailoring feedback. However, current techniques, though effective, often lack interpretability. In this study, we investigate emotion recognition by focusing on making sense of machine-learning models. Specifically, our work leverages the power of interpretable methods in detecting stress through multimodal time series. Building upon prior research, our main contribution is a novel method for calculating feature importance scores using Shapley Additive exPlanations (SHAP) and attention. We applied this technique to models from the MuSe 2022 stress detection competition, generating insights into the importance and interplay of various features in Arousal or Valence prediction. Our findings suggest that leveraging SHAP for feature selection can enhance prediction effectiveness while mitigating computational demands. With this, we introduce an advanced, interpretable paradigm for multi-modal emotion recognition in practical stress-detection scenarios.

Index Terms—Acoustic features, Attention, Emotion (stress) detection, Interpretability, Linguistic features, Multimodal fusion, Multimodal sentiment analysis, Shapley Additive exPlanations (SHAP), Visual features

I. INTRODUCTION

Human emotional responses have long captivated researchers, given their profound influence on our physiology and behavior. Emotional shifts, particularly stress, lead to evident changes in the autonomic nervous system that in turn affect our voice intonation, facial expressions, and body gestures [1]. These are vital cues we innately recognize and interpret. With the advent of artificial intelligence (AI), there has been a surge in attempts to allow machines to recognize these emotional cues, much like humans do [2, 3, 4, 5].

The potential applications of AI-driven emotion recognition are vast, spanning sectors like healthcare, security, and affective computing [6]. Broadly, the techniques

employed fall into two main categories: (1) direct techniques, which require physical contact with the body to obtain readings like heart rate or electroencephalograms [7, 8, 9, 10], and (2) indirect techniques, which rely on cameras and microphones to discern facial gestures, body posture, and speech [11, 12, 13, 14, 15].

While these emotion-recognition techniques have seen notable advancements, they are not without challenges. Many adopt a multimodal approach, extracting information from various sources to enhance their effectiveness. This approach, while improving effectiveness by integrating various modalities [16], complicates the interpretation of how specific modalities impact prediction.

In recent years, the increased effectiveness of machine learning, backed by developments in data engineering and novel learning methodologies, has expanded the horizons of its applications [17, 18, 19]. Such expansion demands not just robust predictions but also a clear understanding of the decision-making process of these models, especially when the outcomes directly impact humans [20]. This is where interpretable machine learning steps in, shedding light on these otherwise ‘black-box’ models [21].

Moreover, in addition to these advancements, it is crucial to discern overarching trends in the decision-making of a model. Understanding these broad patterns helps not only in prediction but also in making sense of the consistency and generalization abilities of models across varied scenarios.

Our foray into this domain began with developing a multimodal Transformer encoder model and engineering a new Pose feature [22]. Building on this foundation, this paper seeks to push the boundaries of emotion recognition through interpretable machine learning. In particular, we aim to:

- Highlight the enhanced effectiveness achievable by fusing various features.
- Perform a detailed investigation of model interpretability, shedding light on why and how specific models learn, by studying Shapley Additive exPlanations (SHAP) and attention scores across a range of features.
- Introduce a novel feature selection methodology that harnesses the power of the SHAP

H. Park, I. G. Kim, J. Oh, and W. De Neve are with the Center for Biosystems and Biotech Data Science, Department of Environmental Technology, Food Technology and Molecular Biotechnology, Ghent University Global Campus, Incheon, Korea. H. Park and W. De Neve are also with IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium.

A. Van Messem is with the Department of Mathematics, University of Liège, Liège, Belgium.

TABLE I: Summary of approaches employed in MuSe-Stress 2022, providing a comparative overview of the various techniques used, including attention mechanisms and LSTM architectures. The acronyms used are as follows: DS denotes DeepSpectrum, eGe denotes eGeMAPS, and ECG denotes Electrocardiogram.

Emotional dimension	Study	Features used	Method	Fusion technique	Test CCC
Arousal	He et al. [23]	DS, IS10, SENetFace, BERTsent, ECG signals	Temporal attention mechanism	-	0.6818
	Liu et al. [24]	DS, FAUs, BERT	LSTM + CNN + self-attention	Early	0.6689
	Li et al. [25]	DS, eGe, ResNet-18, FAUs, BERT, Biosignals	GRU + self-attention	Late	0.5549
	Baseline [15]	DS, eGe, BERT,	LSTM	Late	0.4761
	Ours [22]	DS, eGe, BERT, VGGFace2	Transformer encoder	Late	0.6196
Valence	He et al. [23]	DS, eGe, SENetFace, ECG signals	LSTM	Early	0.6841
	Liu et al. [24]	DS, FAUs, BERT	LSTM + CNN + self-attention	Early	0.6803
	Li et al. [25]	DS, eGe, ResNet-18, FAUs, BERT, Biosignals	GRU + self-attention	Late	0.5857
	Baseline [15]	DS	LSTM	-	0.4931
	Ours [22]	DS, eGe, VGGFace2, FAUs	LSTM	Late	0.6274

scores, demonstrating their utility in handling high-dimensional data.

Our findings point to differences in how features impact Arousal and Valence predictions. Additionally, our findings demonstrate that selecting features based on SHAP scores can enhance prediction accuracy while maintaining computational efficiency. Thus, our findings and the underlying methodology contribute to a better understanding of emotion recognition capabilities in machine learning models and provide insights into how different modalities affect Arousal and Valence predictions.

The remainder of this paper is organized as follows. Section II provides an overview of the different approaches used in the MuSe 2022 Stress Detection competition. Additionally, this section discusses the background and use of interpretable machine learning, which aims to help with the understanding and interpretation of the predictions made. Section III describes the dataset and the methods used for model training and evaluation. Section IV provides an overview of our major experimental results, focusing on unimodal prediction, multimodal late fusion, and conducting an ablation study. Section V analyzes and discusses our results. Finally, Section VI presents conclusions and directions for future research.

II. RELATED WORK

This section discusses our participation in the stress-detection sub-challenge of the 2022 edition of the Multimodal Sentiment Analysis Challenge (MuSe), as well as our application of methods for interpretability. In particular, in Section II-A, we provide an overview of the sub-challenges presented by the MuSe 2022 competition, along with the different methods employed by the participants. Additionally, in Section II-B, we elaborate on SHAP and attention, which we utilized to enable interpretability in our models.

A. MuSe-Stress 2022

MuSe is a renowned emotion recognition challenge. The third edition, *MuSe 2022 Challenge: Multimodal Humour, Emotional Reactions, and Stress*, was held in conjunction with the 30th ACM International Multimedia Conference in Lisbon. This challenge notably featured the MuSe-Stress Sub-challenge, which aimed to assess models capable of predicting emotion labels in 5-minute job interview recordings. MuSe 2022 provided participants with various curated features, such as eGeMAPS from openSMILE [26], facial action units (FAUs) via Py-Feat [27], Biosignals, DeepSpectrum using DenseNet-121 [28], VGGFace2 via ResNet-50 [29], and BERT-based textual features [30]. More details can be found in Section III-B. Furthermore, the organizers set a baseline using a bidirectional long short-term memory (LSTM) model [31] to predict Valence and Arousal from these features, using the Concordance Correlation Coefficient (CCC) as its metric. The MuSe-Stress sub-challenge attracted 41 teams from 16 countries, with four unique methods highlighted in the workshop proceedings. These methods, including ours, are summarized in Table I, along with the baseline approach [15]. In Table I, Test CCC represents the final evaluation result obtained by each team. The detailed metric is discussed in Section III-E.

He et al. [23] extracted features directly, not using the features provided by the organizers of MuSe, except for DeepSpectrum. They re-extracted eGeMAPS and Biosignals from the available raw data by adjusting some of the parameters used by the organizers of MuSe, also adding a new feature called BERTsent, fine-tuned for sentiment analysis tasks in the area of product reviews. Moreover, He et al. proposed a multimodal model using a temporal attention mechanism, training this model with the aforementioned features. The effectiveness of the model obtained was compared with an LSTM model [31] trained using an early fusion approach. In the evaluation on the test set, this LSTM model was able to obtain

the highest effectiveness for predicting Arousal, while also doing well for predicting Valence. Although He et al. [23] acknowledged that their code is based on previous work [32], neither the model nor the parameters used in Muse-Stress have been made publicly available.

Li et al. [25] were active participants in all sub-challenges. For MuSe-Stress, they utilized a ResNet-50 model pre-trained on the VGGFace2 dataset to extract FAUs. They also used the BERT features provided by the organizers, along with re-extracted eGeMAPS, DeepSpectrum, and Biosignals features. Another innovative step they took was to extract a new feature using a ResNet-18 model pre-trained on the mini-AffectNet dataset [33]. For each sub-challenge, the authors developed three distinct models. For MuSe-Stress, in particular, they implemented a Gated Recurrent Unit (GRU) model enhanced with a self-attention mechanism. This model was trained using a combination of DeepSpectrum, FAUs, ResNet-18, and Biosignals features. When evaluating on the test set, a late fusion strategy was employed, with test results surpassing the baseline results provided by the organizers of MuSe. The code and models used by Li et al. are available in their GitHub repository.¹

Among the features made available by the organizers of MuSe, Liu et al. [24] only used DeepSpectrum, FAUs, and BERT. They employed spatio-temporal feature-wise fusion networks, which fed the same inputs into (1) an LSTM and (2) a convolutional neural network with multi-head self-attention. The LSTM output was then used to affine transform the output of (2) using a feature-wise linear modulation method. They employed an ensemble approach, mimicking a late fusion strategy, making it possible to achieve better results than the baseline put forward by the organizers of MuSe. However, neither the model nor the parameters used by Liu et al. in Muse-Stress have been made publicly available.

In summary, most MuSe-Stress participants employed the features provided by the organizers of the challenge, with some of them also adopting a wide variety of alternative feature extraction methods and machine learning techniques, including the attention mechanism [34]. However, it is worth noting that the attention mechanism did not always yield excellent results, and that the use of an LSTM model proved more powerful than expected. In our approach [22], we also made use of an attention-based model, called Transformer encoder, with LSTM performing better in predicting Valence.

None of the papers published in the workshop proceedings, including our paper, adequately explains why and how the underlying approaches work. In this study, we employ interpretability methods to better understand

the input-prediction relationship and to provide more in-depth insights.

B. Interpretability methods

Interpretability in machine learning has become a focal point of research attention, stemming from the growing demand for understanding and trust in models. In this context, we can distinguish between two types of models: white-box models and black-box models, as discussed in [35]. White-box models, exemplified by Decision Trees and Linear Regression, offer a high level of transparency, enabling users to easily understand their inner workings. However, black-box models, which include complex models such as deep neural networks, are usually difficult to interpret directly, necessitating the adoption of specialized methods to understand their decision-making processes, paving the way for further advancements in the field of interpretability.

The concept of interpreting machine learning models and providing insights for human understanding is often referred to as eXplainable Artificial Intelligence (XAI). XAI can be further categorized based on the scope of the interpretation offered. Global interpretation provides a holistic overview, describing the overall behavior of a model, while local interpretation focus on individual predictions, shedding light on the rationale behind specific model decisions [36].

In this section, we focus on the global interpretability of two types of black-box models, namely Transformer encoders and LSTMs, aiming to understand their decision-making processes. We employ tools such as SHAP and the attention score to gain insights into their internal dynamics.

In this study, we utilize SHAP for interpreting the predictions generated by machine learning models, particularly focusing on LSTMs. Based on coalitional game theory [37], this approach is able to estimate the contribution of each input feature to the predictions made by a model. This is achieved by comparing the model output when a feature is present with the model output when the feature is absent. There exist several variants of SHAP, including DeepSHAP, Kernel SHAP [38], and Tree SHAP [39]. Although the basic idea behind these variants is the same, differences can be identified in how they are calculated, depending on the structure of the targeted models. In our research, we utilize Kernel SHAP, which involves the creation of a surrogate model that follows the original model predictions. This surrogate model is trained using a coalition vector that maps the “coalition” of the input features, with the model predictions as output. Then, Shapley values are predicted through various combinations of coalition vectors. Although Kernel SHAP is computationally expensive, it has been demonstrated to have a more stable agreement

¹<https://github.com/MUSE2022-HFUT/MuSe2022>

with human intuition than Local Interpretable Model-Agnostic Explanations (LIME) [40] and Deep Learning Important Features (DeepLIFT) [41], based on human intuition experiments involving 82 random individuals.

We also use the attention mechanism of Transformers [34] as a model-specific method. This mechanism enables a model to focus selectively on parts of the input when processing this input. Attention is typically used in sequence or time series models, such as machine translation and different types of load forecasting, where the input or output can be of different lengths. There is still debate whether attention is related to interpretability [42]. For example, several publications argue that attention is an intermediate result and has little correlation with model interpretation [43, 44], while other papers refute this claim [45, 46]. Nevertheless, attention comes with advantages such as intuition and simplicity, and it has now become a de facto method for interpreting Transformers [47, 48, 49].

III. MATERIALS AND METHODS

This section provides a detailed account of the materials and methods employed in our study. Specifically, we outline the dataset and models utilized, the feature extraction process, the experimental settings, and the employed interpretability methods.

A. Source data: *Ulm-TSST*

The Ulm-Trier Social Stress Test (Ulm-TSST) dataset [15], which was used by the MuSe-Stress 2022 sub-challenge, targets the development of models for predicting continuous emotional Valence and Arousal levels. This dataset includes 69 subjects aged 18 to 39 who were recorded for approximately five minutes in a job interview setting, which is a stress-inducing environment. The total length of the audio-visual recordings available in the Ulm-TSST dataset is 5 hours 47 minutes and 27 seconds. The Rater Aligned Annotation Weighting (RAAW) method [50], involving three raters, has been used to annotate emotional Arousal and Valence levels. In addition to audio-visual recordings, the Ulm-TSST dataset also contains transcripts and physiological signals.

B. Feature extraction

Table II presents the type, name, and dimension of the features used for model training, as well as the extractor used to obtain them. To improve clarity, we divided the features into human-driven features (HDFs) and data-driven features (DDFs). HDFs are features defined by humans, and each element of a feature has a clear and intuitive meaning. For instance, the FAU feature represents a combination of facial muscle movements.

TABLE II: Types of features and their extractors. For clarity, we have categorized the features into two groups: human-driven features (HDFs), which can be easily understood by people, and data-driven features (DDFs), which are extracted using a pre-trained model.

Type	Modal	Name	Dim	Extractor
HDF	Sensor	Biosignals	3	Sampling (2Hz)
	Video	FAUs	20	Py-Feat [27]
	Video	Pose	26	OpenPose [51]
	Audio	eGeMAPS	88	openSMILE [26]
DDF	Video	VGGFace2 [29]	512	ResNet-50 [52]
	Text	BERT	768	BERT [30]
	Audio	DeepSpectrum [28]	1024	DenseNet-121 [53]

On the other hand, DDFs are features extracted using pre-trained convolutional neural networks, making it challenging for humans to interpret the extracted features intuitively.

In what follows, we provide a brief description of each feature type. Before discussing the details of the features used, it is essential to note that, except for the Pose feature, all other features were provided by the organizers of the dataset.

Two acoustic features, eGeMAPS and DeepSpectrum, were used to represent characteristics extracted from the audio modality.

eGeMAPS [54] is a widely-used set of acoustic features in audio processing and analysis. It consists of 88 low-level acoustic HDF features that combine low-level descriptors and functionalities. The low-level descriptors are values extracted from a speech signal, which include pitch, jitter, shimmer, spectral flux, spectral roll-off, spectral energy, loudness, and voicing probability. The functionalities are methods used to represent each feature, such as first and second-order derivatives, arithmetic mean, and standard deviation. The extraction of eGeMAPS features was facilitated through the use of the openSMILE toolkit.²

DeepSpectrum refers to a collection of audio features derived from Mel spectrograms. These features are extracted by a pre-trained DenseNet-121 network [28]. The Mel spectrograms are utilized as a tool to depict the spectral content of an audio signal. This feature was extracted through code from the eponymous Github repository.³

BERT pertains to features that have been extracted from German textual transcripts by employing the BERT model [30], while VGGFace2 and FAUs refer to features that have been extracted from the video modality.

The VGGFace2 features were extracted by utilizing a ResNet-50 network that has been pre-trained on the VGGFace2 dataset [29], which comprises approximately

²<https://github.com/audereing/opensmile>

³<https://github.com/DeepSpectrum/DeepSpectrum>

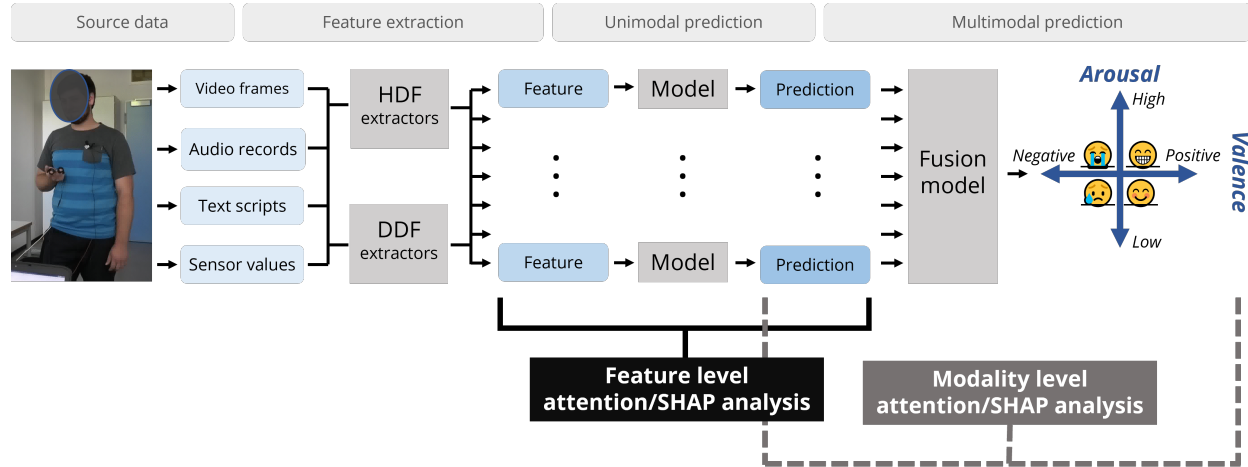


Fig. 1: Overall pipeline for multimodal stress detection: (1) feature extraction, (2) unimodal prediction, and (3) multimodal prediction. The resulting model can predict the level of emotional Arousal or Valence. In this study, we investigate which uni/multimodal features have the greatest impact on our best models, hereby also paying attention to the influence of their feature dimensions.

3.3 million facial images of approximately 9,000 distinct individuals.

FAUs [55] pertain to a distinct facial region that can aid in predicting the type and intensity of emotions being felt. The Py-Feat [27] tool (default configuration) was used to extract FAU features.

Biosignals refers to a combination of three sensor values: electrocardiogram (ECG), respiration (resp), and heart rate (BPM). We treated these three values as a single modality, consistent with the approach taken in the MuSe-Stress 2022 baseline.

The Pose feature was extracted by utilizing the OpenPose GitHub project,⁴ which tracks the movement of 25 key points of the human body across several video frames.

C. Models

In this study, we utilized two models: the first being the LSTM model provided by the MuSe-Stress organizers [15], and the second being the Transformer encoder model we introduced in previous work [22]. In Table I, two approaches obtained higher CCC values than our approach. However, for these two approaches, neither the code nor the model parameters used for the competition could be obtained from the authors. Therefore, among the different approaches published, we chose our model, which has the highest CCC value, as our main model. We employed late fusion, as depicted in Figure 1. In the unimodal prediction step, we trained a separate model for each unimodal feature and emotional dimension. Subsequently, in the multimodal step, we used as input a

combination of the predictions produced by the unimodal models. At each stage, each model predicted only one emotional dimension, either Arousal or Valence. We trained twenty identical models using different random seeds and we fine-tuned the hyperparameters using a two-pronged approach [22]. Specifically, this approach involved shallow tuning of model capacity, the number of stacks, and learning rates, and deep tuning of window lengths, hop lengths, and batch sizes. After performing shallow and deep tuning, the final test submission was obtained by adopting a late fusion approach, selecting deep-tuned unimodal models with high Development CCC scores.

The prediction outcomes of the two models utilized in our earlier study are shown in Table III. The highest Development/Test CCC scores for each emotional dimension, modality, and type of model are emphasized in bold. In addition, we conducted an in-depth analysis of these models using SHAP and attention scores.

In Table III, the two emotional dimensions, namely Arousal and Valence, are both on display with the results obtained by the LSTM model and the Transformer encoder model listed for each dimension. To further understand the decision-making process of these models, we employed SHAP scores with the LSTM model and attention scores with the Transformer encoder model. Detailed explanations are provided in Section III-F. The test set labels were withheld (that is, the test set labels were not publicly available), and the evaluation on the test set was carried out through an online platform provided by the organizers of MuSe-Stress. The column with the label Devel CCC represents the Development CCC scores obtained for the publicly available valida-

⁴<https://github.com/CMU-Perceptual-Computing-Lab/openpose>

TABLE III: A summary of the prediction outcomes obtained in previous work [22]. A, T, and V are abbreviations used in late fusion, where A refers to DeepSpectrum and eGeMAPS, T refers to BERT, and finally, V refers to VGGFace2 and FAUs. The Test CCC results are largely empty because of a submission limit. The highest Devel/Test CCC scores for each emotional dimension, modality, and type of model are emphasized in bold. This study employs SHAP and attention to investigate the models that generated the prediction outcomes shown below.

Features		(Physiological) Arousal				Valence			
		LSTM		Transformer encoder		LSTM		Transformer encoder	
		Devel CCC	Test CCC	Devel CCC	Test CCC	Devel CCC	Test CCC	Devel CCC	Test CCC
Unimodal features	DeepSpectrum	0.4880	0.4220	0.4521	-	0.6140	-	0.6424	0.5581
	eGeMAPS	0.5238	-	0.6635	0.4191	0.6100	0.3722	0.5901	-
	BERT	0.4915	-	0.5503	0.3504	0.5518	0.3974	0.4470	-
	Biosignals	0.5310	0.1477	0.3399	-	0.4629	0.1127	0.3503	-
	VGGFace2	0.5876	0.2342	0.4576	-	0.4319	0.1362	0.4902	-
	FAUs	0.6127	0.0043	0.5296	-	0.5543	0.4459	0.4529	-
	Pose	0.5216	-	0.5454	0.1764	0.5047	-	0.6025	0.2079
Late fusion	A+T	0.6818	-	0.6810	-	0.6652	-	0.6810	-
	A+V	0.6598	-	0.7660	-	0.8096	0.6281	0.6642	-
	T+V	0.5989	-	0.6400	-	0.7061	-	0.7450	-
	A+T+V	0.6787	-	0.7560	-	0.7291	-	0.7620	-
	Best 3 max	0.7660	-	0.5872	-	0.6528	-	0.7840	-
	Best 3 mean	0.7773	-	0.7820	-	0.7760	-	0.6634	-
	Total	0.7369	-	0.7940	0.4591	0.8080	-	0.6990	-
Final	DS+eGe+BERT+VGG	-	-	0.6973	0.6196	-	-	-	-
	A+V	-	-	-	-	0.8125	0.6351	-	-

tion (development) set. When compared to the baseline, except for the Arousal-DeepSpectrum-Test CCC, we can observe that both the Devel and Test CCC of our LSTM and Transformer encoder exhibit superior performance. The CCC results for the baseline and comparisons with our results can be found in the baseline paper [15] and our previous work [22], respectively. Due to the limited number of submissions available for testing, the Test CCC portion of Table III is largely unpopulated. Detailed explanations of Table III and the test submission strategy used in previous research are available in Appendix D. Test CCC scores and the methodologies used to derive them are discussed in [22] and can also be found on our GitHub page.⁵

D. Hardware and software settings

Our experiments were carried out employing three NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory, and two Intel(R) Xeon(R) Silver 4214R CPUs (2.40 GHz, 12 cores, and 24 threads). We used Python 3.8.8, PyTorch 1.11.0, and CUDA 11.3 to train our models and perform hyperparameter tuning. We also utilized Weights and Biases 0.12.16 for logging and tracking. To avoid overfitting, we used the default early stop strategy of the MuSe-Stress 2022 baseline code during training. Additionally, we trained the models for 100 epochs. Typically, training the LSTM baseline model used approximately 1,500 to 3,000 MiB of GPU memory, while training our Transformer encoder utilized around 3,000 to 5,000 MiB of GPU memory.

⁵<https://github.com/powersimmani/MuSe2022FeelsGood>

E. Evaluation metric and loss

We used the CCC as an evaluation metric and as a loss to train both the LSTM baseline model and our Transformer encoder. The CCC is defined as:

$$CCC = \frac{2\rho\sigma_{\hat{Y}}\sigma_Y}{\sigma_{\hat{Y}}^2 + \sigma_Y^2 + (\mu_{\hat{Y}} - \mu_Y)^2},$$

where μ_Z and σ_Z are the mean and the standard deviation of $Z \in \{Y, \hat{Y}\}$, and where ρ is the Pearson correlation coefficient between \hat{Y} (prediction) and Y (label). Furthermore, the CCC-based loss \mathcal{L} is given as follows:

$$\mathcal{L} = 1 - CCC.$$

Note that CCC is a statistical measure that evaluates the degree of agreement between predicted and actual values. Although the Pearson correlation coefficient could be used for this purpose, it only measures the strength of a linear relationship. For instance, if \hat{y}_i are exactly twice the y_i , the Pearson correlation between Y and \hat{Y} would be 1, but the difference between the predicted and actual values would be large. The key difference in CCC comes from changing the denominator (penalty term) from $\sigma_{\hat{Y}} \cdot \sigma_Y$ to $\sigma_{\hat{Y}}^2 + \sigma_Y^2 + (\mu_{\hat{Y}} - \mu_Y)^2$. This modification means that CCC penalizes for the variability and the difference in means between the observed and predicted values, allowing for more accurate measurement of the agreement between the ground truth and predictions.

F. Interpretability methods

In this section, we introduce two approaches for assessing the importance of the unimodal and multimodal features utilized by our models: the SHAP score and the attention score. The SHAP score is computed using the Kernel SHAP value, while the attention score is the sum of the attention output weights. We also contrast the dissimilarities between these two approaches and describe how each one is integrated into the model.

1) *SHAP score*: Let n be the length of the time series and m be the dimension of each feature (see “Dim” in Table II). Furthermore, for $i = 1, \dots, m$, let \mathbf{x}_i be the i th observation of $X \in \mathbb{R}^{m \times n}$. In addition, let $\hat{\mathbf{y}}$ be the vector of length n with the prediction outcomes. Suppose we have an optimized stress detection model $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^n$ that has already been trained:

$$\hat{\mathbf{y}} = f(X) \in \mathbb{R}^n.$$

Given the trained model f and the dataset X , for $j = 1, \dots, n$, we can use the Kernel SHAP approach [38] to approximate $f(\mathbf{x}_j)$ with $g(\mathbf{z}'_j)$ as follows:

$$g(\mathbf{z}'_j) = \phi_{j0} + \sum_{i=1}^m \phi_{ji} z'_i, \quad (1)$$

where ϕ_{j0} is the bias, $\mathbf{z}' \in \{0, 1\}^m$ is a simplified input that is working as a coalition vector, and $\phi_{ji} \in \mathbb{R}$ ($i = 1, \dots, m; j = 1, \dots, n$) is a feature attribution. For example, ϕ_{ji} represents the contribution of the i th feature of \mathbf{x}_j to obtain the prediction $f(\mathbf{x}_j)$.

To focus on the feature attribution ϕ , let $\Phi(\mathbf{x}_j)$ denote the feature attribution vector of the input \mathbf{x}_j for the j th time step, i.e.,

$$\Phi(\mathbf{x}_j) = [\phi_{j1} \quad \dots \quad \phi_{jm}].$$

We define the SHAP score vector $\mathbf{s}_{\text{score}} \in \mathbb{R}^m$ using Φ as follows:

$$\mathbf{s}_{\text{score}} = \left[\sum_{j=1}^n |\phi_{j1}| \quad \dots \quad \sum_{j=1}^n |\phi_{jm}| \right]. \quad (2)$$

Note that each element of $\mathbf{s}_{\text{score}}$ denotes the overall feature importance, which is computed by taking the absolute sum of the feature importance at each time point.

2) *Attention score*: We define the single-head attention weight matrix $\mathbf{H} \in \mathbb{R}^{n \times n}$ as follows:

$$\mathbf{H}(Q, K) = \text{Softmax} \left(\frac{Q \cdot K^\top + M}{\sqrt{d_h}} \right), \quad (3)$$

where n is the length of the sequence, d_h is the embedding dimension of the attention layer of a Transformer encoder, $Q \in \mathbb{R}^{n \times d_h}$ is the query, $K \in \mathbb{R}^{n \times d_h}$ is the

key, and $M \in \mathbb{R}^{n \times n}$ is the mask (that is, the triangular matrix in the attention layer). The attention weights form a square matrix whose dimension equals the sequence length.

Assuming that we have q heads, the mean attention weight matrix \mathbf{B} is then defined as follows:

$$\mathbf{B} = \frac{1}{q} \sum_{h=1}^q \mathbf{H}_h(Q, K) \in \mathbb{R}^{n \times n}. \quad (4)$$

Finally, let us denote the elements of \mathbf{B} corresponding to the i th row and the j th column as \mathbf{B}_{ij} . The attention score vector $\mathbf{a}_{\text{score}} \in \mathbb{R}^n$ can then be defined as follows:

$$\mathbf{a}_{\text{score}} = \left[\sum_{i=1}^n \mathbf{B}_{i1} \quad \dots \quad \sum_{i=1}^n \mathbf{B}_{in} \right]. \quad (5)$$

Due to the usage of Softmax, the sum of the entries of $\mathbf{a}_{\text{score}}$ is n .

3) *Method differences*: The first method employed is SHAP, which generates, for each input vector, coalition vectors representing possible combinations of features. The SHAP algorithm subsequently calculates the contribution of each feature to the model output by simulating the model behavior on possible coalitions of features. This results in an overview of which features are deemed important. However, it is worth noting that SHAP scores are calculated and summarized for each input vector independently, which means it does not account for temporal aspects, such as the significance of specific time points throughout an entire video.

In contrast, the attention-based method is particularly useful for analyzing time series as it identifies essential time points or sections of the input for decision-making. Nevertheless, the embedding of all features in the Transformer encoder complicates pinpointing the overall important features.

In Table IV, we summarize a number of key differences between the models and the interpretability methods used. In conclusion, both methods have their own strengths and weaknesses, and can be used together in a complementary manner for uni/multimodal time series prediction. In this study, we utilized SHAP scoring with the LSTM model and attention scoring with the Transformer encoder. This approach was adopted as the results indicated a lack of clear superiority or inferiority between the LSTM and Transformer encoder models. Additionally, the models operate differently, making it unfeasible to apply attention with LSTM and SHAP with the Transformer encoder.

Finally, our rationale for choosing a global approach like SHAP over other local XAI techniques is as follows: in the context of model interpretation, while SHAP gives a holistic view of the behavior of a model, local explanations, such as counterfactual explanations [56],









Remark	Highly ranked in both Arousal and Valence		Highly ranked in Arousal Lowly ranked in Valence			Lowly ranked in Arousal Highly ranked in Valence		
Description	AU 05  Upper lid raiser	AU 10  Upper lip raiser	AU 04  Brow lowerer	AU 06  Cheek raiser	AU 07  Lid tightener	AU 12  Lip corner puller	AU 24  Lip pressor	AU 25  Lips part
Top 5 in Arousal	✓	✓	✓	✓	✓	□	□	□
Top 5 in Valence	✓	✓	□	□	□	✓	✓	✓

Fig. 2: Overview of highly ranked FAU features (top 5) and lowly ranked FAU features (not top 5) according to s_{score} . We can observe that some FAU features exhibit bias towards an emotional dimension (black vs dark grey).

TABLE IV: Overview of the models used in our study, alongside the corresponding methods and the specific dimensions of data that these methods can interpret.

Our model	Interpretability method	Interpretability dimension
Transformer encoder	Attention score (a_{score})	Time series (n)
LSTM	SHAP score (s_{score})	Feature dimension (m)

emphasize individual predictions. Considering our objective to grasp overarching patterns in the decision-making of a model, the use of a global approach like SHAP was deemed more fitting than the use of intricate, instance-focused counterfactual explanations.

IV. RESULTS

In what follows, we present our experimental results. In Section IV-A, we detail the results obtained for unimodal features, particularly for HDF, where the number of features is relatively small, and where the results are easy to comprehend. In Section IV-B, we present the outcomes obtained for multimodal fusion. Lastly, in Section IV-C, we introduce and discuss a feature selection method that discards features with a low SHAP score.

A. Score analysis for unimodal prediction

1) *SHAP score*: In this section, we concentrate on the SHAP score results related to unimodal features, especially emphasizing FAUs, eGeMAPS, and Pose, which are classified as HDFs. We deliberately left out Biosignals and DDFs from our discussion, because both in Arousal and Valence analyses, Biosignals did not reach a Test CCC score higher than 0.15, which consequently led to fusion disregarding the inclusion of Biosignals. On the other hand, while DDFs played a role in late fusion, we chose not to cover them in this section due to their high dimensionality complicating a straightforward

interpretation. However, for readers interested in a more in-depth exploration, the detailed discussions and raw SHAP scores for Biosignals and DDFs are available in Appendix A, with a comprehensive analysis presented in Appendix D.

Figure 2 compares the ranking of FAU features over emotional dimensions using their s_{score} values. Features ranked within the top 5 are considered highly ranked, while those outside this range are deemed lowly ranked. FAU features colored in black and dark gray signify features that are highly ranked in one emotional dimension but lowly ranked in the other emotional dimension. For instance, our experimental results indicate the following:

- both AU05 and AU10 rank within the top 5 for both Arousal and Valence;
- AU04, AU06, and AU07 are highly ranked for Arousal, but they are lowly ranked for Valence;
- AU12, AU24, and AU25 are highly ranked for Valence, but they are lowly ranked for Arousal.

These results suggest that the influence of a feature can differ based on the emotional dimension being predicted. It should be noted that Figure 2 was adapted from an image by Zhi et al. [55].

Table V provides a summary of eGeMAPS features with high s_{score} values, yielding models that obtained a top-2 or top-3 rank in terms of both Devel and Test CCC (amongst models only making use of unimodal features). Our experimental results show that low-level descriptors related to frequency, such as Formant $F1$, Formant $F2$, and Formant $F3$ frequencies [54, 57, 58], semitone frequency [57, 59], and the second and third coefficients of Mel-Frequency Cepstral Coefficients (MFCC) [60], have a high impact. On the other hand, energy/amplitude-related parameters, such as loudness or spectral (balance) parameters, including the Alpha Ratio [54], exhibit little importance. Since the dataset is composed of interviews, we can deduce that the model focuses on the vocalizations of the subjects. Moreover, unlike FAUs, there are fewer rank differences between Arousal and Valence.

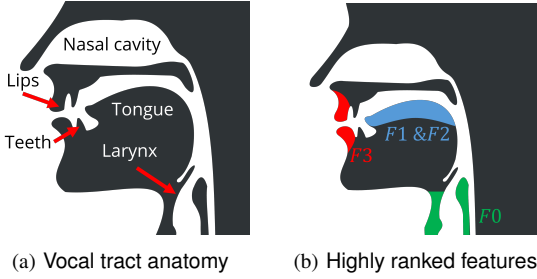


Fig. 3: Diagram of the vocal tract anatomy. The highly ranked eGeMAPS features listed in Table V are closely related to parts that produce human voice and speech, such as the tongue (blue), lips (red), and larynx (green). $F0$ refers to the fundamental frequency and $F1$ to $F3$ refer to the Formant, one of the low-level descriptors of eGeMAPS. Each formant is marked in the color corresponding to the associated anatomical part.

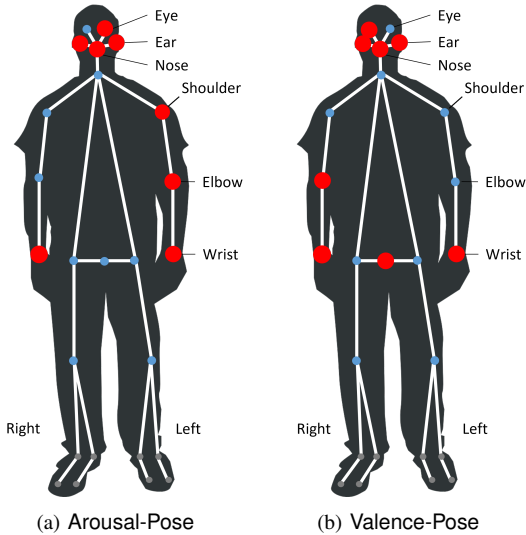


Fig. 4: The human body plots show the importance of different Pose features based on s_{score} . The red dots indicate the body parts that are considered critical by the model, with bigger sizes indicating higher importance. Substantial differences between the Arousal and Valence models were observed in the elbow and eye regions.

To enhance the comprehension of our research findings, Figure 3 displays the vocal tract anatomy, linking the eGeMAPS features from Table V to their anatomical origins:

- Figure 3(a) presents the primary components.
- Figure 3(b) highlights the top-ranked features using:
 - Blue for the tongue, influencing $F1$ and $F2$.
 - Red for the lips, impacting $F3$.
 - Green for the larynx, linked to $F0$.

Figure 4 presents the importance of different Pose

features, based on s_{score} values obtained for the most effective LSTM models trained on the unimodal Pose feature.

Typically, body parts that are barely visible due to the position of the camera, such as the feet and knees, have an s_{score} value of zero. The hands, which demonstrate the most movement, obtain the highest s_{score} values. Interestingly, we can observe that the Arousal and Valence prediction models emphasize different body parts. Specifically, the Arousal prediction model prioritizes the left elbow, shoulder, and eye, while the Valence prediction model prioritizes the right elbow and eye.

2) *Attention score*: The proposed attention score allows us to identify high-attention time intervals in the Transformer encoder, thereby improving model interpretability. However, unlike the original Transformer model, which represents a single interval as an embedded single word [34], our Transformer encoder is trained on a dataset in which each time interval is represented as a feature with dimensions ranging from 3 to 1024, which required us to find a way to interpret these high-dimensional results. To that end, we used a method that compares and divides the data into, on the one hand, time intervals that received a top-1% attention score and, on the other hand, the other time intervals as described below.

First, we used t-distributed Stochastic Neighbor Embedding (t-SNE) [61] to reduce the dimensionality of each HDF and DDF feature. We then segmented the results into 20 clusters using K -means clustering [62]. Subsequently, we identified the cluster that received the top-1% a_{score} and named it the Most Significant Cluster (MSC). We then compared the MSC with the other clusters to comprehend the characteristics of the features within the MSC. Our objective was to investigate whether there was a noticeable difference in value distribution between MSC and non-MSC features with high SHAP scores. However, upon closer examination, we found no clear connection between SHAP scores and attention scores.

Nevertheless, we found a notable difference for the FAU features, as depicted in Figure 5. For instance, Figure 5(a) illustrates the Transformer encoder model trained with Arousal-FAUs, where the orange box represents the distribution of feature values in the MSC, and the blue box represents the feature value distribution of all other clusters. We discovered that the higher the attention received, the higher the values of AU10, AU06, AU07, AU12, and AU25. In contrast, for Valence-FAUs, the differences were smaller, but we observed that the feature values of AU04, AU06, and AU07 in the MSC were, on average, slightly higher than those in the non-MSC clusters.

Moreover, we observed variations in the distribution

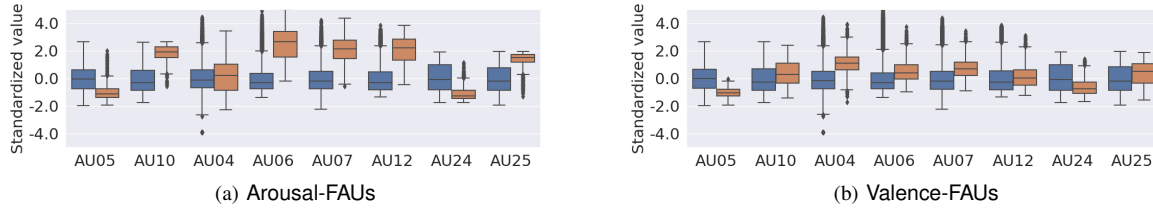


Fig. 5: Comparison of the distribution of FAU feature values. The distribution of the feature values in the MSC is highlighted in orange, while the distribution of the feature values in the remaining clusters is highlighted in blue. The order of the features is based on Figure 2.

TABLE V: Highly ranked eGeMAPS features according to s_{score} values. These features are crucial for predicting both Arousal and Valence.

Importance	Low-level descriptors	Functionalities	Related to
Top 5	Formant F3 frequency and bandwidth	Arithmetic mean	Vowel, lip shape, roundness shape
	Formant F2 frequency and bandwidth	Arithmetic mean	Vowel, tongue position, length of lingualis anterior
	Formant F1 bandwidth	Arithmetic mean	Vowel, tongue position, volume of the pharyngeal cavity
Top 10	Formant F1 frequency	Arithmetic mean	Vowel, tongue position, volume of the pharyngeal cavity
	F0 semitone frequency from 27.5Hz	Stddev, rising slope	Intonation, frequency of laryngeal vibrations
	2nd and 3rd coefficient of MFCC	Normalized stddev	Spectrum, shape and centroid of the sound spectrum

of feature values between MSCs and non-MSCs that did not correspond to the features deemed important by SHAP scoring. As this information is too extensive to be included in the main text, we make the results available in Appendix D for readers interested in further analysis.

B. Score analysis for multimodal late fusion

In this section, we present an analysis of SHAP scoring and attention scoring for the two types of multimodal late fusion models used to generate the final Test CCC values. The obtained results can be found in Figure 6. For Arousal, the Transformer encoder was trained using late fusion of the DeepSpectrum, eGeMAPS, BERT, and VGGFace2 predictions. For Valence, the LSTM was trained using late fusion of the DeepSpectrum, eGeMAPS, FAU, and VGGFace2 predictions. In the case of the Transformer encoder (Arousal), we reduced the dimension of the input features to two dimensions using t-SNE, leading to easier interpretability. We marked the attention scores in the top 1% (i.e., larger or equal than the 99th percentile) in orange. As can be seen in Figure 6(a), these dots are concentrated on one side.

We can compare the feature value distribution by making a distinction between the features that fall into the top-1% attention scores (orange) and all other features (blue), as shown in Figure 6(b). For each of the four features — DeepSpectrum, eGeMAPS, BERT, and VGGFace2 — the median value of the orange box is higher than that of the blue box. This observation could imply that, within each feature, higher values (in this case, unimodal predictions) tend to receive greater attention. However, it is important to note that the

differences between the orange and blue boxes, except for eGeMAPS, do not seem highly significant.

In the case of LSTM (Valence), SHAP scoring was used to determine the importance of each feature. The obtained results are shown in Figure 6(c). The features illustrated in the plot are eGeMAPS (0.37), DeepSpectrum (0.32), FAUs (0.28), and VGGFace2 (0.27). The obtained scores are all relatively high compared to a baseline of zero, which would indicate no feature importance. This suggests that all four features play a considerable role in the prediction of Valence by the LSTM model. However, it is worth noting that the audio-type features (eGeMAPS and DeepSpectrum) have slightly higher SHAP values compared to the visual-type features (FAUs and VGGFace2), indicating that the audio modality might be relatively more important for predicting Valence in this particular model.

C. Feature selection using SHAP scores

During the process of obtaining SHAP and attention scores, we could observe an interesting pattern: certain features received very low a_{score} values. For instance, as can be seen in the leftmost part of Figure 1, the knee of the subject was omitted. As a result, for the Pose feature, key points of the knee and key points below the knee are likely to receive lower SHAP scores compared to other key points, as confirmed in Figure 4. We also noticed this pattern for eGeMAPS, DeepSpectrum, and VGGFace2. Given these observations, we propose the use of a feature selection method that leverages SHAP scores to make a distinction between important and unimportant features, where the latter can be discarded.

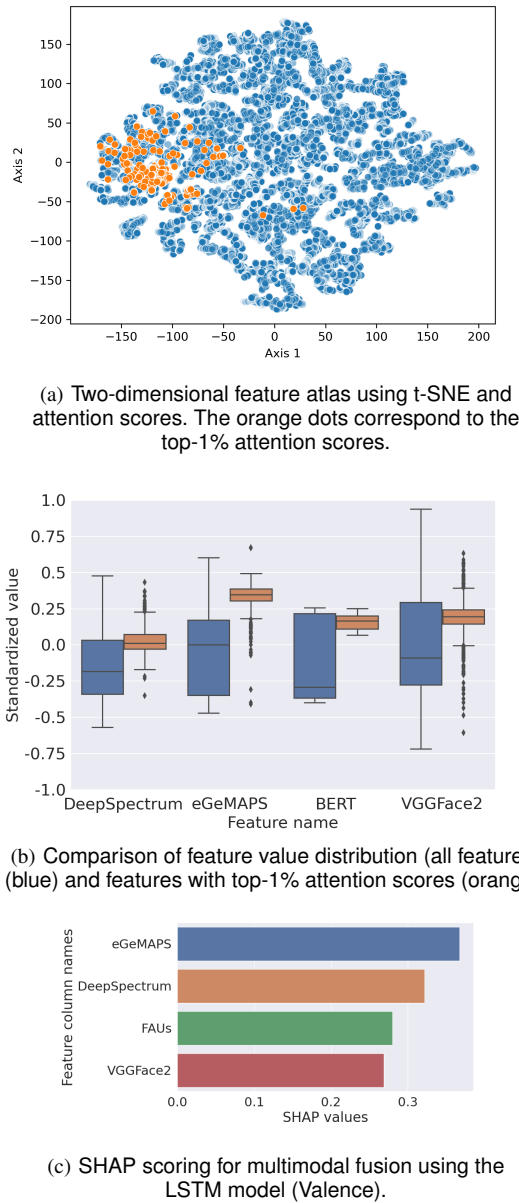


Fig. 6: Results obtained by our final late fusion models. The scatter plot (a) and the box plot (b) show the results obtained by the Transformer encoder model when predicting Arousal, whereas the bar plot (c) shows the obtained SHAP scores for the LSTM model. The orange color in (a) and (b) represents the input features that got higher (99th percentile or above) attention scores, while the blue color represents all other input features.

Figure 7 presents the outcomes of our ablation study, illustrating the effectiveness of the model as we sequentially remove features with the smallest SHAP scores. This experiment was conducted by bringing together

the two emotional dimensions and three unimodal features, resulting in six combinations. The solid blue line represents the Devel CCC, whereas the orange dashed line denotes the Test CCC. The x -axis of each subplot indicates the number of features, hereby starting from the left with all features (88, 512, or 1024) and removing features when moving to the right. The Test CCC was calculated using additional submission attempts provided by the MuSe-Stress submission system after the end of the competition.⁶

Our results demonstrate that using 78 features in Arousal-eGeMAPS scores better than the use of the original 88 features in terms of Test CCC, whereas the use of 38 features in Valence-eGeMAPS results in an optimal Test CCC. In Arousal-VGGFace2, we can observe that the Test CCC decreases when the number of features decreases to 404. However, further reducing the number of features to 284 leads to an improvement in both Test CCC and Devel CCC. Using 404 features in Valence-VGGFace2 results in a substantial increase in Test CCC. In Arousal-DeepSpectrum, the Test CCC improves as the number of features decreases. However, in Valence-DeepSpectrum, there is a slight decline in both Test and Devel CCC. Overall, the Test CCC improves for four out of six combinations by reducing the number of features. More detailed investigations can be found in Appendix D.

V. DISCUSSION

The study presented in this paper leveraged the Ulm-TSST dataset, which embodies both multimodal and time series characteristics, to discern stress indicators. Our previous work [22] resulted in the third place in the MuSe-Stress 2022 challenge. Whereas prior research focused on improving effectiveness using a novel Pose feature and a Transformer encoder, the research effort presented in this paper aims at improving our understanding of the importance of features using SHAP and attention, extracted from models trained on Ulm-TSST.

A. Key findings

One important finding from our study is the variability in feature significance. Investigating FAUs and Pose provided insightful examples, emphasizing the asymmetric value of specific body parts and FAUs in predicting Arousal and Valence. This observation points to potential undiscovered correlations between emotional and physiological responses. Our analysis further demonstrated that eGeMAPS predominantly focuses on human speech features rather than generic sound features.

⁶https://codalab.lisn.upsaclay.fr/competitions/5357/?secret_key=832c58ee-e9e1-4e36-9b82-e6819767f43f

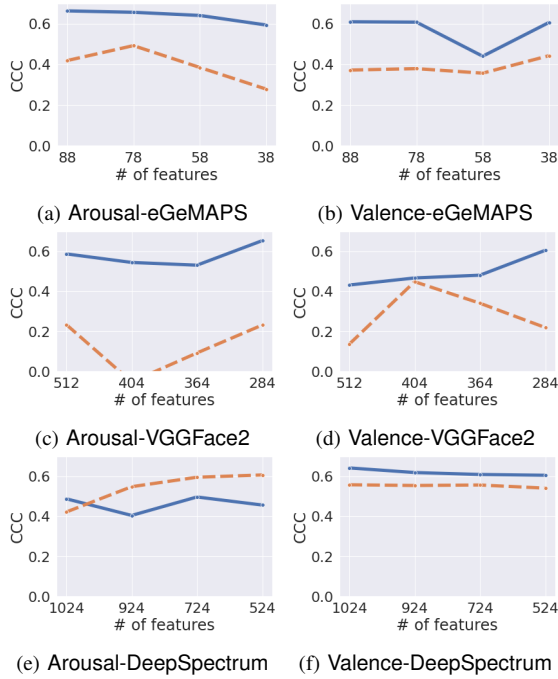


Fig. 7: Results obtained by our ablation study, showing the effect of removing features, in order of smallest SHAP score, on model effectiveness. The solid blue lines represent the Devel CCC, whereas the dashed orange lines represent the Test CCC.

When employing multimodal models, we noticed a distinct preference for audio features, indicating that the fusion model gives precedence to eGeMAPS and DeepSpectrum features. The utilization of SHAP for feature selection played a fundamental role in our study, underscoring its effectiveness in both reducing computational overhead and being useful even when new DDFs are introduced in the future.

B. Limitations of SHAP in audio feature interpretability

SHAP values, grounded in feature independence and symmetry, may face challenges when applied to features that do not exhibit the two aforementioned properties. For example, the position of the tongue during speech impacts Formant $F1$ and $F2$, illustrating feature interdependencies. This complexity can be easily found in audio-type features like eGeMAPS and DeepSpectrum. Consequently, relying solely on the foundational assumptions of SHAP might not capture the full understanding of audio data, suggesting the need for further research on tailored interpretability techniques or hybrid methods for more accurate analysis.

C. The scope and shortcomings of the Ulm-TSST dataset

The Ulm-TSST dataset distinctively enhances sentiment analysis with its extensive scope, surpassing similar datasets like IEMOCAP [11], MSP-IMPROV [13], and RAVDESS [14], by tripling the subject pool. Its uniqueness stems from capturing authentic emotional responses in a real-world scenario, unlike others that primarily use actor-simulated emotions. This characteristic enriches its practical applicability, offering a more natural perspective on emotional analysis.

However, the widespread applicability of multimodal sentiment analysis must take into account certain inherent limitations. A prime example is the demographic composition of the Ulm-TSST dataset. Dominated by German speakers aged between 18 and 39 years, there exists a risk of it not entirely capturing a varied demographic, which could introduce potential biases, thereby challenging the generalizability of the results derived from it.

VI. CONCLUSIONS

In this study, we focused on applying interpretability in sentiment analysis problems characterized by multimodality and time series properties. We investigated models for sentiment analysis with the aim of achieving a better understanding of the underlying reasons and mechanisms behind their decision-making processes, thus going beyond merely evaluating their effectiveness. Our study revealed several insights, such as the differential importance of each HDF feature across Arousal and Valence, and the feasibility of reducing computational demands while maintaining effectiveness through the application of s_{score} for feature selection.

The primary insights of this study are derived exclusively from the Ulm-TSST dataset. To enhance the generalizability of our findings, future research efforts should consider applying our approach to a broader spectrum of datasets such as EMOCAP, MSP-IMPROV, RAVDESS, and even newly constructed datasets that cover a large number of subjects, a wide range of (genuine) emotional responses, and more use cases. Additionally, exploring alternative methods for interpretability, especially in scenarios where the assumptions of SHAP regarding feature independence and symmetry are less applicable might be crucial. Methods like Integrated Gradients [63] could be less influenced by these two assumptions. On the other hand, conducting local case studies using counterfactual explanations or LIME could also provide further valuable insights. These methods show how small input changes can make a big difference in predictions, helping us understand how models make decisions.

Furthermore, investigating ways to leverage insights from different emotion recognition tasks to develop more

robust and generalizable models is a promising direction. This could involve studying how interpretability methods can help identify common patterns across tasks such as stress detection, humor recognition, and reaction intensity prediction, all of which were part of MuSe-2022. Such cross-task analysis could potentially lead to more universal observations about emotion recognition models and their interpretability. Finally, designing new interpretability methods specifically for datasets with multimodal and time series characteristics could also offer novel insights and contribute substantially to the field of affective computing, particularly in areas such as stress detection and sentiment analysis. Additionally, expanding on the ablation analysis in Section IV-C to uncover the reasons behind each observation could provide valuable insights into model behavior and feature importance.

While our study provides valuable insights, it is important to acknowledge the limitations of the Ulm-TSST dataset, particularly in terms of demographic representation. Future research should aim to incorporate more diverse datasets to enhance the generalizability of findings in multimodal sentiment analysis. Additionally, as we continue to develop and apply these technologies, it is crucial to maintain a focus on fairness and ethics. Interpretability methods play a key role in this regard, helping to identify and mitigate potential biases and ensuring the responsible deployment of emotion recognition systems across various domains.

In conclusion, this research effort represents a step toward addressing the following question in the practical application of sentiment analysis models: "Are they trustworthy?" Through the application of methods for interpretability, we improve not only model transparency but also facilitate more reliable and ethically sound applications of sentiment analysis in diverse situations.

APPENDIX A SHAP SCORES

This appendix contains all information used to obtain the SHAP scores presented in Section IV-A1. The files make use of the CSV format.

Link: <https://figshare.com/s/8bae7b015a0054a5de08>

APPENDIX B ATTENTION SCORES

This appendix contains all information used to obtain the attention scores presented in Section IV-A2. The files make use of the Python Pickle format.

Link: <https://figshare.com/s/8bae7b015a0054a5de08>

APPENDIX C SELECTED FEATURES

This appendix contains the selected features used for training, as discussed in Section IV-C.

Link: <https://figshare.com/s/0f725364144ba22cdefc>

APPENDIX D SUPPLEMENTARY MATERIALS

This appendix provides in-depth analyses that could not be included in the main text due to space constraints. Specifically, it includes an in-depth analysis of SHAP scores and attention scores, with detailed visualizations for both HDFs and DDFs. Additionally, it offers further analysis of feature selection using SHAP scores, as well as additional explanatory notes for Table III and the test submission strategy employed in previous research. Link: <https://figshare.com/s/9495ee00db5456bda578>

REFERENCES

- [1] S. D. Kreibitz, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, 2010, the biopsychology of emotion: Current theoretical and empirical perspectives. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0301051110000827>
- [2] J. Zhang, H. Yin, J. Zhang, G. Yang, J. Qin, and L. He, "Real-time mental stress detection using multimodality expressions with a deep learning framework," *Frontiers in Neuroscience*, vol. 16, 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fnins.2022.947168>
- [3] R. Li and Z. Liu, "Stress detection using deep neural networks," *BMC Medical Informatics and Decision Making*, vol. 20, no. 11, p. 285, Dec 2020. [Online]. Available: <https://doi.org/10.1186/s12911-020-01299-4>
- [4] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," 2022. [Online]. Available: <https://arxiv.org/abs/2203.07378>
- [5] S. Siriwardhana, T. Kaluarachchi, M. Billingham, and S. Nanayakkara, "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.
- [6] A. I. Middy, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950068722001085>

- ect.com/science/article/pii/S0950705122002593
- [7] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 81–84.
- [8] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.
- [9] S. Katsigiannis and N. Ramzan, "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2018.
- [10] K. A. Araño, P. Gloor, C. Orsenigo, and C. Vercellis, "Emotions are the Great Captains of Our Lives": Measuring Moods Through the Power of Physiological and Environmental Sensing," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1378–1389, 2022.
- [11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec 2008. [Online]. Available: <https://doi.org/10.1007/s10579-008-9076-6>
- [12] S. Parthasarathy, C. Zhang, J. H. Hansen, and C. Busso, "A study of speaker verification performance with expressive speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5540–5544.
- [13] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.
- [14] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018. [Online]. Available: <https://doi.org/10.1371/journal.pone.0196391>
- [15] L. Christ, S. Amiriparian, A. Baird, P. Tzirakis, A. Kathan, N. Mueller, L. Stappen, E. Messner, A. König, A. Cowen, E. Cambria, and B. Schuller, "The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress," 04 2022.
- [16] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Meßner, E. Cambria, G. Zhao, and B. W. Schuller, "The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, ser. MuSe '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 5–14. [Online]. Available: <https://doi.org/10.1145/3475957.3484450>
- [17] C. C. S. Liem, M. Langer, A. Demetriou, A. M. F. Hiemstra, A. Sukma Wicaksana, M. P. Born, and C. J. König, *Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening*. Cham: Springer International Publishing, 2018, pp. 197–253. [Online]. Available: https://doi.org/10.1007/978-3-319-98131-4_9
- [18] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.
- [19] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.
- [20] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, Oct. 2017. [Online]. Available: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2741>
- [21] C. Molnar, *Interpretable machine learning*. Lulu.com, 2020.
- [22] H.-m. Park, I. Yun, A. Kumar, A. K. Singh, B. J. Choi, D. Singh, and W. De Neve, "Towards Multimodal Prediction of Time-Continuous Emotion Using Pose Feature Engineering and a Transformer Encoder," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, ser. MuSe' 22. New York, NY, USA: Association for Computing Machinery, 2022, p. 47–54. [Online]. Available: <https://doi.org/10.1145/3551876.3554807>
- [23] Y. He, L. Sun, Z. Lian, B. Liu, J. Tao, M. Wang, and Y. Cheng, "Multimodal Temporal Attention in Sentiment Analysis," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, ser. MuSe' 22. New York, NY, USA: Association for Computing Machinery, 2022, p. 61–66. [Online]. Available: <https://doi.org/10.1145/3551876.3554811>

- [24] Y. Liu, W. Sun, X. Zhang, and Y. Qin, "Improving Dimensional Emotion Recognition via Feature-Wise Fusion," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, ser. MuSe' 22. New York, NY, USA: Association for Computing Machinery, 2022, p. 55–60. [Online]. Available: <https://doi.org/10.1145/3551876.3554804>
- [25] J. Li, Z. Zhang, J. Lang, Y. Jiang, L. An, P. Zou, Y. Xu, S. Gao, J. Lin, C. Fan, X. Sun, and M. Wang, "Hybrid Multimodal Feature Extraction, Mining and Fusion for Sentiment Analysis," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, ser. MuSe' 22. New York, NY, USA: Association for Computing Machinery, 2022, p. 81–88. [Online]. Available: <https://doi.org/10.1145/3551876.3554809>
- [26] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: <https://doi.org/10.1145/1873951.1874246>
- [27] E. Jolly, J. H. Cheong, T. Xie, S. Byrne, M. Kenny, and L. J. Chang, "Py-Feat: Python Facial Expression Analysis Toolbox," 2021. [Online]. Available: <https://arxiv.org/abs/2104.03509>
- [28] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 26–29.
- [29] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 67–74.
- [30] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [32] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2020.
- [33] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," in *IEEE Transactions on Affective Computing*. IEEE, 2017.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- [35] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019. [Online]. Available: <https://doi.org/10.1038/s42256-019-0048-x>
- [36] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," 2017.
- [37] L. S. Shapley, *Notes on the N-Person Game — II: The Value of an N-Person Game*. Santa Monica, CA: RAND Corporation, 1951.
- [38] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>
- [39] S. M. Lundberg, G. G. Erion, and S. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," *CoRR*, vol. abs/1802.03888, 2018. [Online]. Available: <http://arxiv.org/abs/1802.03888>
- [40] M. T. Ribeiro, S. Singh, and 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*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [41] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3145–3153.
- [42] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, and P. Watrin, "Is Attention Explanation? An Introduction to the

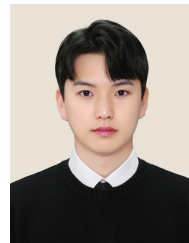
- Debate,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3889–3900. [Online]. Available: <https://aclanthology.org/2022.acl-long.269>
- [43] S. Jain and B. C. Wallace, “Attention is not Explanation,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3543–3556. [Online]. Available: <https://aclanthology.org/N19-1357>
- [44] B. Bai, J. Liang, G. Zhang, H. Li, K. Bai, and F. Wang, “Why Attentions May Not Be Interpretable?” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 25–34. [Online]. Available: <https://doi.org/10.1145/3447548.3467307>
- [45] S. Wiegrefe and Y. Pinter, “Attention is not not Explanation,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. [Online]. Available: <https://aclanthology.org/D19-1002>
- [46] A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran, “Towards Transparent and Explainable Attention Models,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4206–4216. [Online]. Available: <https://aclanthology.org/2020.acl-main.387>
- [47] H. Chefer, S. Gur, and L. Wolf, “Transformer Interpretability Beyond Attention Visualization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 782–791.
- [48] J. Vig, “A Multiscale Visualization of Attention in the Transformer Model,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 37–42. [Online]. Available: <https://aclanthology.org/P19-3007>
- [49] E. Gholami, M. Motamedi, and A. Aravindakshan, “PARSRec: Explainable Personalized Attention-Fused Recurrent Sequential Recommendation Using Session Partial Actions,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 454–464. [Online]. Available: <https://doi.org/10.1145/3534678.3539432>
- [50] L. Stappen, L. Schumann, B. Sertolli, A. Baird, B. Weigell, E. Cambria, and B. W. Schuller, “MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox,” in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, ser. MuSe ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 75–82. [Online]. Available: <https://doi.org/10.1145/3475957.3484451>
- [51] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [52] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.
- [54] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
- [55] R. Zhi, M. Liu, and D. Zhang, “A comprehensive survey on automatic facial action unit analysis,” *The Visual Computer*, vol. 36, no. 5, pp. 1067–1093, May 2020. [Online]. Available: <https://doi.org/10.1007/s00371-019-01707-5>
- [56] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR,” *Harvard Journal of Law & Technology (Harvard JOLT)*, vol. 31, no. 2, pp. 841–888, 2018 2017. [Online]. Available: <https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=859>
- [57] S. A. Memon, “Acoustic Correlates of the Voice Qualifiers: A Survey,” *CoRR*, vol. abs/2010.15869, 2020. [Online]. Available: <https://arxiv.org/abs/2010.15869>

10.15869

- [58] S. Heo and H. Kang, "Formant frequency changes of female voice /a/, /i/, /u/ in real ear," *Phonetics and Speech Sciences*, vol. 9, no. 1, pp. 49–53, 2017. [Online]. Available: <http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07133926>
- [59] J. O. Uguru, "Fundamental frequency as cue to intonation: Focus on Ika Igbo and English rising intonation patterns," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3573–3573, 2013. [Online]. Available: <https://doi.org/10.1121/1.4806554>
- [60] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Chapter 3 - Features for Content-Based Audio Retrieval," in *Advances in Computers: Improving the Web*, ser. Advances in Computers. Elsevier, 2010, vol. 78, pp. 71–150. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0065245810780037>
- [61] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: <http://jmlr.org/papers/v9/vandermaaten08a.html>
- [62] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [63] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3319–3328. [Online]. Available: <https://proceedings.mlr.press/v70/sundararajan17a.html>



Ganghyun Kim Ganghyun Kim is an undergraduate student at Ghent University Global Campus (GUGC) in the Republic of Korea, pursuing a bachelor's degree in Molecular Biotechnology. Presently, he undertakes a role as a research intern at the Center for Biosystems and Biotech Data Science. His areas of research interest encompass computer vision and reinforcement learning.



Jinsung Oh Jinsung Oh is an undergraduate student at Ghent University Global Campus (GUGC) in the Republic of Korea, pursuing a bachelor's degree in Molecular Biotechnology. Currently, he holds a position as a research intern at the Center for Biosystems and Biotech Data Science, focusing on applying machine learning and computer vision techniques to medical imaging and industrial sensor data.



Arnout Van Messem Arnout Van Messem earned degrees in Mathematics and Actuarial Sciences from the Vrije Universiteit Brussel, Belgium. In 2011, he completed his PhD at the same university. Presently, he holds a professorship at the Department of Mathematics of the Université de Liège in Belgium. His research focuses on machine learning, non-parametric statistics, missing data, and robustness.



Ho-min Park Ho-min Park earned his MSE degree in Computer Engineering from Ajou University, Republic of Korea, in 2018. Following this, he began his doctoral studies in Computer Science Engineering at Ghent University, Belgium. Currently, he serves as a research and teaching assistant at Ghent University Global Campus (GUGC), Republic of Korea. His research focuses on applied computer vision and machine learning, exploring their applications in environmental monitoring, medical imaging, and industrial sensor signal processing.



Wesley De Neve Wesley De Neve earned his M.Sc. degree in Computer Science and a Ph.D. in Computer Science Engineering from Ghent University, Belgium, in 2002 and 2007, respectively. Currently, he serves as an Associate Professor at both IDLab, Ghent University, Belgium, and the Center for Biosystems and Biotech Data Science at Ghent University Global Campus (GUGC) in the Republic of Korea. His primary research focus lies in utilizing (deep) machine learning to extract knowledge from biological sequences and biomedical imagery.