Multimodal Prediction of Alexithymia from Physiological and Audio Signals

Valeria Filippou
CASTORC
The Cyprus Institute
Nicosia, Cyprus
v.filippou@cyi.ac.cy

Mihalis A. Nicolaou
CASTORC
The Cyprus Institute
Nicosia, Cyprus
m.nicolaou@cyi.ac.cy

Nikolas Theodosiou
CASTORC
The Cyprus Institute
Nicosia, Cyprus
n.theodosiou@cyi.ac.cy

Georgia Panayiotou
Department of Psychology
The University of Cyprus
Nicosia, Cyprus
georgiap@ucy.ac.cy

Elena Contantinou
Department of Psychology
The University of Cyprus
Nicosia, Cyprus
constantinou.c.elena@ucy.ac.cy

Marios Theodorou
Department of Psychology
The University of Cyprus
Nicosia, Cyprus
theodorou.marios@ucy.ac.cy

Maria Panteli
Department of Psychology
The University of Cyprus
Nicosia, Cyprus
mariapant@windowslive.com

***Abstract*—Alexithymia is a trait that reflects a person's difficulty in recognising and expressing their emotions, which has been associated with various forms of mental illness. Identifying alexithymia can have therapeutic, preventive, and diagnostic benefits. However, there has been limited research on proposing predictive models for alexithymia, and literature on multimodal approaches is almost non-existent. In this light, we present a novel predictive framework that utilises multimodal physiological and audio signals, such as heart rate, skin conductance level, facial electromyograms, and speech recordings to detect and classify alexithymia. To this end, two novel datasets were collected through an emotion processing imagery experiment, and subsequently utilised on the task of alexithymia classification by adopting the TAS-20 (Toronto Alexithymia Scale). Furthermore, we developed a set of temporal features that both capture spectral information and are localised in the time-domain (e.g., via wavelets). Using the extracted features, simple machine learning classifiers can be used in the proposed framework, achieving up to 96% f1-score - even when using data from only one of the 12 stages of the experiment. Interestingly, we also find that combining auditory and physiological features in a multimodal manner further improves classification outcomes. The datasets are made available on request by following the provided github link.**

***Index Terms*—Affective Computing, Multimodal Machine Learning, Alexithymia.**

I. Introduction

Alexithymia, a personality trait, was originally defined by Sifneos in 1972 [1] as the difficulty in recognising, naming, and describing emotions with words. Alexithymic individuals tend to focus their thoughts on external stimuli [2]. Although it is not considered a psychiatric disorder, it is associated with various mental and physical health issues [3]. Clinically significant alexithymia affects 10% of the general population [4].

The Toronto Alexithymia Scale (TAS-20) is a popular self-reporting tool used to subjectively evaluate alexithymia. Since it is the self-reporting instrument that is most frequently utilised, it is regarded as the industry standard [5]. However, it has been highlighted that alexithymia is also characterised by variances in physiological responsiveness to emotional experiences, despite the fact that research in the subject has frequently produced conflicting conclusions about the nature of these distinctions [6]. Additionally, these physiological variations are frequently believed to be key elements contributing to the phenomenology of alexithymia, as well as its causation and maintenance, although this assumption [7] has not yet been scientifically supported. The core difficulties of alexithymia must be identified in order to determine the relative contribution of physiological markers during emotion processing. To do so, researchers can utilise cutting-edge technological and statistical methods, such as Machine Learning (ML), which enables the simultaneous assessment of multiple signals as they unfold dynamically over time [8]. Additionally, effective predictive models have the potential to be implemented in embedded systems, notably wearables. Technical advantages include shorter experimental times, real-time remote patient monitoring, less intrusive physiological signal measurement, and cost-effectiveness compared to more expensive ML algorithms [9]. Indeed, this line of research can help us better understand alexithymic deficits, such as whether they are arousal-related (depending on the intensity of emotions as reflected in Autonomic Nervous System (ANS) signals like skin conductance) or valence-related (depending on the unpleasantness of a situation as reflected in facial expression signals). Importantly, it can also have clinical implications since detecting and treating alexithymic issues could stop the emergence of diseases that are closely associated to them in the future (i.e. depression).

To the best of our knowledge, this is one of the *very few* attempts to exploit multi-modal signals as predictors for alexithymia using ML. We provide a method for detecting alexithymia using five physiological signals that were obtained from four different modalities (electrocardiogram (ECG), electrodermal activity (EDA), and electromyography (EMG) [10]) and speech. We specifically design a novel set of features that

capture spectral and temporal information from multimodal signals, and then cross-modal correlations in the feature space are captured using simple classification techniques. We assess our approach using two datasets from studies that employed emotional imagery to evoke emotions with the goal of using the TAS-20 to spot variations between control and alexithymic volunteers [11]. We demonstrate that the suggested framework can result in an f1-score of up to 96% using data from only a tiny section of the experiment, a finding that can result in a reduction in the amount of time and money spent on the experiment. The efficiency of the suggested method may make it appropriate for low-power and embedded systems, and it may also increase how well self-reporting tools are used.

This study expands on previous research [11]–[13] by (1) including more subjects, (2) an additional modality (audio), (3) new experiments using pre-trained models for audio, as well as (4) new experiments and findings based on multimodal fusion.

We take into account the challenge of creating reliable alexithymia baseline classifiers given five time-series physiological and audio data from two datasets. Effective generalisation is hindered by three factors: a lack of large open-access datasets, a lack of established protocols that adhere to best standards for evaluation, and excessive subject variability.

Our study recruited over 100 participants to gather physiological and auditory data for creating accurate classification models for alexithymia, a novel subject requiring more data. The absence of a uniform evaluation process presents an obstacle to advancement, as various papers may not use the same experimental layout, rendering results incomparable. Subject variability, including sensor positioning, head and body motion, and noise, presents challenges for generalising to new subjects or sessions. Without open data and benchmarks, comparing distinct models is difficult.

The remainder of this paper is structured as follows: In Section II, we overview related literature, contrasting to this work. In Sections III, IV, and V we describe the datasets, methods, and results of the study respectively, while Section VI discusses findings, limitations, and future research directions.

II. Related Work

Emotion recognition using physiological signals has gained significant attention from researchers in recent years [14]. Several studies have been conducted to explore the potential of physiological signals, such as electroencephalogram (EEG), electrocardiogram (ECG), and skin conductance response (SCR), in recognising human emotions. The use of ML algorithms to analyse physiological signals has enabled researchers to develop more accurate and robust emotion recognition systems. In this section, we present some of the notable works that have been done in this area [15].

There are two main approaches to ML: classical ML algorithms that involve manual feature engineering (also known as hand-crafting), and Deep Learning (DL) models that learn hierarchical, compositional representations for specific tasks. Various feature engineering techniques are available to extract and select useful features. Some studies have used sequential forward and backward feature selection methods [16], while others have manually extracted features [9], [17]. Furthermore, these studies have found that between five and 14 features are needed to achieve high performance scores for emotion recognition [9], [16], [17].

In recent years, researchers have turned to DL algorithms to predict emotions and their components, such as arousal, valence, and dominance. The most commonly used DL algorithms are Deep Neural Network (DNN) [18], [19], Convolutional Neural Network (CNN) [18]–[20], and recurrent neural network (RNN) models like Long Short-Term Memory (LSTM) [19], [21], [22]. For example, in one study, DNN and CNN models were developed, and in another study [21], a bimodal LSTM was created for emotion recognition. Both studies achieved mean accuracy greater than 75% for the classification of valence and arousal. Other studies, [20] developed models for anxiety detection using different physiological signals, such as using a 1D CNN trained on ECG-based features to detect anxiety in arachnophobic individuals, achieving an accuracy of 83.29%. Generally, DL models outperformed simpler, traditional algorithms because they can better optimise extracted features [19], [22].

In addition, some studies have used spectral features extracted from physiological signals, which were then converted into images and used in pre-trained models for classification tasks. For example, both [23] and [24] used deep transfer learning for emotion recognition using physiological signals to classify arousal and valence. Only one paper, Nima et al. [8], was identified as predicting alexithymia using ML, but by recording a video of volunteers' facial expressions.

This paper contributes to the field of affective computing by: (1) introducing a model that predicts alexithymia using physiological and audio signals, including ANS indices of arousal and facial responses to emotional valence; (2) using the largest dataset available for classifying alexithymia with physiological signals; (3) identifying fft_coefficients and cwt_coefficients as descriptive features for alexithymia classification; (4) conducting extensive experiments to analyse the effect of hyperparameters on model performance, and (5) making available upon request the two datasets of physiological signals.

III. Datasets

A. Subjects

Young and healthy adults from two universities in Cyprus participated in this study. The Greek-translated TAS-20 was used to screen them for alexithymia, which is assessed through difficulty identifying feelings, difficulty describing feelings, and externally oriented thinking on a 5-point scale. The total score was used to determine clinical levels of alexithymia, with scores of 51 and below indicating low alexithymia, 52-59 indicating medium, and 60 and above indicating high. Exclusion criteria included medical or mental health conditions that could affect ANS reactivity and regular medication use [13].

For Dataset 1, we used a total of 54 eligible participants who were categorised into two groups based on their alexithymia

levels: 27 participants with high alexithymia and 27 with low alexithymia. The average age of the participants was 21.36 (SD = 2.95). Among the participants, eight identified as male, while 42 identified as female. Half of the male participants (n=4) were categorised as having low alexithymia, while the other half (n=4) were categorised as having high alexithymia. Among female participants, 21 were categorised as having low alexithymia, and 21 were categorised as having high alexithymia. We are missing demographic data from four volunteers.

For Dataset 2, we included 65 eligible participants who were categorised into three groups based on their alexithymia levels: 11 participants with high alexithymia, 10 with medium alexithymia, and 44 with low alexithymia. The participants had an average age of 21.18 (SD = 2.52). Among the participants, 16 identified as male, while 48 identified as female. Of the female participants, eight were categorised as having high alexithymia, eight as having medium alexithymia, and 32 as having low alexithymia. Among the male participants, one was categorised as having medium alexithymia, three as having high alexithymia, and 12 as having low alexithymia. We are missing demographic data from one volunteer.

B. Experiment design

In this experiment, participants were provided with ten standardised emotional scripts. The scripts were divided into two categories of depth of processing: shallow processing and deep processing [13]. The instructions for shallow and deep processing differed in terms of the aspects of imagery to be emphasised. For shallow processing, participants were asked to vividly visualise the scene while recalling the details of the imagined locations, such as objects, people, and animals. For deep processing, they were directed to focus on affective reactions and subjective experiences, visualising the scene as if they were actively participating.

Each participant performed five scripts under shallow processing conditions and five scripts under deep processing conditions. Both depths of processing followed the same procedure for each script, which involved a rest period of 20 seconds, followed by a 60-second phase 1 and a 40-second phase 2.

C. Data acquisition

1) Data sources: BIOPAC MP150 for Windows and AcqKnowledge 3.9.0 data acquisition software (Biopac Systems Inc., Santa Barbara, CA) were used to obtain physiological data. The researcher placed electrodes on the arms and face of the participants following standard procedures. Additionally, a tablet was used to record the audio recordings for dataset 2 solely.

The dependent variables included two signals that reflect levels of arousal (heart rate (HR) and skin conductance level (SCL)), and three signals that reflect changes in experienced valence (orbicularis (ORB), corrugator (COR), zygomaticus (ZYG)) measured during both baseline and emotional imagery. At the end of each imagery trial, participants provided self-reports of how they felt, including emotional labeling, valence, arousal, and dominance ratings. The physiological signals were filtered as described in detail by Constantinou et al. [11]. In addition, for Dataset 2, participants provided audio recordings expressing how they felt.

2) Imagery materials: Ten standardised emotional scripts were selected from a larger pool of validated emotional scripts that were specific to the community. The ten scripts represented typical fear, joy, and neutral scenarios. Joy and fear scripts were selected to differ considerably on valence but not on arousal, while neutral scripts differed on both dimensions from both joy and fear scripts. Affective imagery has been successfully used to induce emotions of varied valence and arousal levels [25]. These three types of emotions allowed for the independent investigation of the effects of valence and arousal on physiological reactivity. The scripts were provided to all participants in three semi-counterbalanced orders. They were written in the first person, two sentences long, and included references to physical reactions. Further information and examples of the scripts can be found in Constantinou et al. [11].

3) Experimental protocol and set-up: Upon arrival at the lab, participants were seated in a reclining chair placed in a dark, sound-proof room. After providing informed consent, they were given instructions and fitted with physiological monitors. Prior to the experiment, a five-minute adjustment phase was conducted to stabilise physiological data and familiarise participants with the apparatus. Next, participants were instructed on the depth of processing required for the trial, and given an index card containing the imagery script to memorise. At the sound of a tone, they began visualising the scripts. No individuals were excluded from the study for not completing the task, and there were no differences in task performance across all groups [11].

IV. Methods

Deep Learning (DL) is commonly applied nowadays since the developed network can learn and decide effectively on its own. However, because the dimensionality of the signals is frequently greater than the number of individuals (under the same experimental settings), DL findings may be subpar. The results of the pilot study [12] confirmed this. In this light, we present a set of spectral features for alexithymia detection using statistical hypothesis time-series on multiple time-series features. In this method, we decrease the dimensionality of the problem to a few scalars per time-series signal. Our findings show that the developed features are discriminative for alexithymia, with f1-score of up to 96% when only one stage of the trial, e.g. Fear-Deep-Phase-1 (FDP1) or Joy-Deep-Phase-2 (JDP2), is used.

A. Data pre-processing

The BIOPAC software files were imported into PythonTM (v.3.8.8) for analysis. The extracted files had many columns that represented time, HR, SCL, ORB, COR, ZYG, and digital

channels denoting the experiment's stage. The digital channels (i.e. experiment stages) were phase-1, phase-2, arousal, valence, tone-1 (shallow processing), and tone-2 (deep processing), all of which were represented by binary representations. The phases were utilised to denote the experiment's imagery period, which were: baseline (20sec), first imagery period (60sec), and second imagery period (40sec). The terms arousal and valence were employed to describe the emotion of the script. Finally, the tones were utilised to indicate whether the experiment's processing level was shallow or deep. As previously stated, this describes how individuals reacted to what they imagined at the end of the experiment.

Table I defines the signals and their modalities.

TABLE I: Modalities used to measure physiological signals

Modality	**Definition**	**Physiological signal**
ECG	Measures the potential differences identified at the skin surface due to electrical activity of the heart	HR
EDA	Measures the electrical conductivity of the skin	SCL
EMG	Measures the skeletal muscle electric activity at the skin surface. It is used for both facial or body expressions	ZYG, COR, ORB
Speech	Represents the acoustic properties of the spoken language	Audio

All of the physiological signals were captured at a sampling frequency of 1000 Hz[1]. The following step was to select only the data of interest, which contained the data from the ten trials. This means that any data from inter-trial intervals was eliminated. The values of the digital signals indicated this, and the data was divided into separate stages depending on these values. Additionally, all the audio signals were recorded with a sampling frequency of 16000 Hz.

The raw data was independently downsampled for each participant to improve memory complexity. The polyphase filter resampling [26] with down-sampling factor N = 300 was used to achieve the down-sampling. This number is equal to three seconds, which is regarded sufficient for identifying emotional shifts (empirical and trial-based evidence) [27]. The data was resized to meet the needs of each algorithm.

B. Feature extraction and selection

Physiological signals and audio signals have distinct characteristics. Heart rate or skin conductance, may require specific feature selection methods tailored to their unique properties. Similarly, audio signals may have different spectral, temporal, or perceptual features that necessitate specific approaches for feature selection. Using modality-specific feature selection methods allows for better capturing and representing the relevant information present in each modality. Hence, tsfresh[2] was used for physiological signals, and librosa[3] was used for audio signals.

1) Physiological signals: The tsfresh library was used to efficiently extract and select relevant features from multivariate signals [28]. Tsfresh extracts (794) time-series characteristics in order to explain a time series dataset in relation to a target variable. Statistical hypothesis testing is used to assess the discriminative performance and significance of retrieved characteristics for a certain task. In our example, we chose the ten most relevant attributes to provide into the classifier.

Several tsfresh traits were retrieved for each individual experiment, resulting in 3915 attributes for each subject. The importance of the aforementioned qualities was computed using the target variable in order to keep the top ten for each subject. As a result, the final data structure was a matrix X $\in \mathbb{R}^{n\times 10}$, where n is the number of participants. Table II shows the top 10 features for the universal models of both datasets 1 and 2, where for dataset 2 only low and high alexithymic participants were included.

TABLE II: Top k=10 features extracted via tsfresh. (*a*: Variance, *b*: Absolute differences, *c*: Higher quantile, *d*: Lower quantile, *e*: Imaginary, *f*: Absolute)

Dataset 1 - Universal
FFT Coefficient (Real), Coeff 20, ORB
Change Quantiles, F_agg Var[a], Isabs[b] False, Qh[c] 0.4, Ql[d] 0.0, ZYG
Change Quantiles, F_agg Var, Isabs True, Qh 0.4, Ql 0.0, ZYG
FFT Coefficient (Imag[e]), Coeff 50, ECG
FFT Coefficient (Imag), Coeff 40, SCR
FFT Coefficient (Abs[f]), Coeff 33, SCR
FFT Coefficient (Abs), Coeff 69, ORB
FFT Coefficient (Real), Coeff 68, COR
Change Quantiles, F_agg Mean, Isabs True, Qh 0.4, Ql 0.0, ZYG
Change Quantiles, F_agg Var, Isabs True, Qh 0.2, Ql 0.0, ZYG
Dataset 2 - Universal
FFT Coefficient (Real), Coeff 50, SCR
FFT Coefficient (Angle), Coeff 50, SCR
CWT Coefficients, Coeff 4, W 2, Widths (2, 5, 10, 20), ZYG
FFT Coefficient (Imag), Coeff 5, SCR
FFT Coefficient (Imag), Coeff 5, ORB
FFT Coefficient (Real), Coeff 10, ZYG
FFT Coefficient (Angle), Coeff 80, SCR
FFT Coefficient (Real), Coeff 10, COR
CWT Coefficients, Coeff 10, W 2, Widths (2, 5, 10, 20), SCR
CWT Coefficients, Coeff 1, W 5, Widths (2, 5, 10, 20), ZYG

After looking at the top characteristics from all the results, the most common features that occurred were `fft_coefficient` and `cwt_coefficients`. The former, in particular, computes the discrete Fourier Transform coefficients [29], which are provided by:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{i2\pi}{N} kn} \tag{1}$$

where N is the total number of samples in the input sequence.

[1]Except for phase-2, which had a sampling frequency of 125 Hz in some cases due to human error. As a result, the initial step was to resample phase-2 signals to 1000 Hz as needed.

[2]https://tsfresh.readthedocs.io/en/latest/

[3]https://librosa.org/doc/latest/index.html

The Fast Fourier Transform technique is used [30]. The coefficient is calculated for either the real, imaginary, magnitude, or angle in degrees' components of the expansion. `Cwt_coefficients`, on the other hand, provides a Continuous Wavelet Transform for the 'Mexican hat wavelet' [31] provided by:

$$\psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1 - \frac{t^2}{\sigma^2}\right)e^{\frac{-t^2}{2\sigma^2}} \quad (2)$$

where σ is the scale factor. When compared to other wavelets, empirical evidence reveals that the specific wavelet may characterise a signal with a very minimal number of parameters [32]. The aforementioned features are retrieved using a parallel feature selection approach that is based on statistical hypothesis tests like the Mann-Whitney U [33] or the Kolmogorov Smirnov [34]. These tests are set up based on the label type (categorical or continuous) and the supervised ML task at hand (regression or classification).

2) Audio signals: Many features were calculated from audio sources using the librosa library. The nine Mel Frequency Cepstral Coefficients (MFCC), zero-crossing rate, spectral roll-off, spectral centroid, spectral contrast, spectral bandwidth, and delta of MFCC coefficients were determined [35]. After that, the retrieved features are concatenated into a matrix. The following features are available for each type of feature extraction: mean, standard deviation, skewness, maximum, median, and minimum values. The final matrix had 90 features for each participant.

A DT was used to find the most descriptive features, and then the top 10 features were selected to be used as input data into the ML algorithms.

C. Classification

Deep neural networks (DNNs), decision trees (DT), random forests (RF), multilayer percpetrons (MLP), and logistic regression (LR) models with various hyper-parameter configurations were used for the classification models.

1) Deep Neural Networks: Training DL models from scratch is a time-consuming and data-intensive procedure. In the pilot study [12], DNNs based on pre-trained networks such as ResNet [36], DenseNet [37] and AlexNet [38] all of which were pre-trained on the Image-Net dataset [39], did not perform well on our dataset since the data we have differs significantly from the data the pre-trained models were trained with. However, in this study we used Wav2Vec [40] and HuBERT [41] pre-trained models on the audio dataset.

2) Logistic Regression: The LR algorithm is a classification algorithm. Based on a set of independent variables, it is used to calculate (or forecast) a binary (yes/no) event. The model creates a regression model to predict the likelihood that a given data input belongs to the '1' category. The sigmoid function is used by LR to model the data. ($g(z) = \frac{1}{1+e^{-z}}$).

3) Decision Tree: This is a tree-based method in which nodes represent features, leaves represent outcomes, and branches represent decisions. The DT method divides the dataset into smaller subsets depending on features until all of the sample points have a final label. The algorithm uses gini impurity to choose the optimal split, beginning with the root node and on to the subsequent splits, ($Gini = 1 - \sum_{i=1}^{c} p_i^2$), where p_i is the probability of the class *i* in a node. Gini impurity chooses the best possible split by measuring the split's quality. The impurity with the lowest and best value is zero. When all of the samples have the same label, this is achieved.

4) Random Forest: A parallel ensemble method is the RF algorithm. Ensemble methods are a class of techniques that combine numerous ML algorithms into a single predictive model. This is done to reduce bias (boosting), variation (bagging), or to improve predictions (stacking). In greater detail, boosting aims to reduce bias by iteratively adjusting the weights or focus on misclassified instances. Bagging helps to decrease variation by aggregating predictions from multiple models trained on different subsets of the data. Stacking leverages the predictions of multiple models as input to a meta-model, which improves overall prediction accuracy. RF is a DT ensemble, which means that it constructs numerous DTs and combines them through a voting procedure to get a more stable and accurate forecast. RF belongs to the bagging algorithm family.

5) Multilayer Perceptron: MLPs are neural networks composed of interconnected nodes (perceptrons) that use weighted inputs and an activation function to process data. They are capable of learning complex patterns in non-linearly separable data through iterative training using backpropagation.

D. Evaluation

The models were evaluated using leave-one-subject-out cross-validation, a form of *k*-fold cross-validation where *k* equals the number of participants in the dataset. This strategy ensures subject-independence, utilises most of the information, and prevents over-fitting.

V. Results

Two datasets were used to classify the level of alexithymia. Dataset 1 included subjects with either low or high alexithymia. Dataset 2 on the other hand included subjects with low, medium, or high alexithymia. Furthermore, physiological and audio signals were obtained for dataset 2.

Figure 1 shows the physiological data distribution for the two datasets with only low and high alexithymic participants. To visualise the distribution of the five physiological signals a boxplot was used for each signal.

The project involved testing several instances, including binary classification of physiological signals, multi-class classification of physiological signals, and multi-class classification of both physiological and audio signals. For the *binary classification* scenario of physiological signals, **four** distinct input data scenarios were used: features generated using the **tsfresh** library from i) dataset 1, ii) dataset 2, iii) dataset 1 & 2, and iv) concatenated top 10 features from dataset 1 and dataset 2 separately. The data used in this case included individuals with either **low or high** alexithymia levels. For the

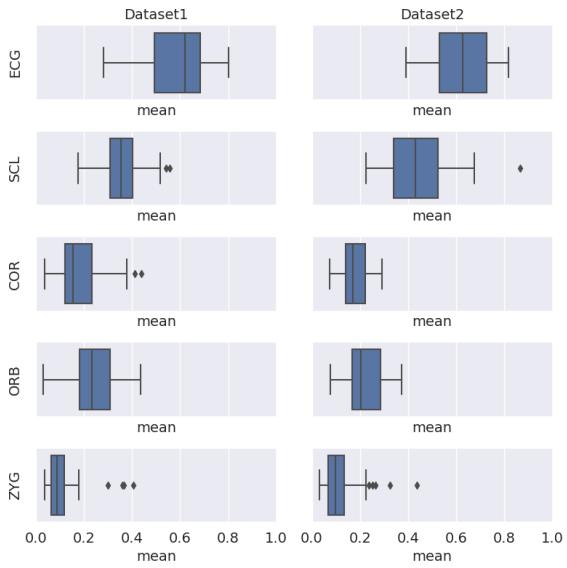

Fig. 1: Exploratory data analysis of datasets 1 & 2

multi-class classification scenario of physiological signals, **two** input data scenarios were used: features generated using the **tsfresh** library from i) dataset 2, and ii) dataset 1 & 2. The data used in this case included individuals with **low, medium, or high** alexithymia levels. In the *multi-class classification* of both physiological and audio signals scenario, data exclusively from **dataset 2** was used, which included individuals with low, medium, or high alexithymia levels. **Four** distinct input data scenarios were used: i) the original raw audio signals were used as input in two pre-trained models, ii) extracted audio features using librosa library, iii) extracted physiological features using tsfresh library, and iv) the top 10 audio features concatenated with the top 10 physiological signals.

A. *Physiological signals - Binary classification*

Table III displays the results of several experiments that correspond to specific stages of the trial. The experiments involved using tsfresh features that were calculated from different combinations of `dataset 1` and `dataset 2`, as well as concatenated features from both datasets. Four ML algorithms were utilised to classify participants as either *low* or *high* alexithymics. LR produced the highest f1-score, achieving 96% in fear-shallow-phase2 dataset 2. RF followed closely with 94% f1-score for fear-deep-phase1 dataset 1. MLP achieved an f1-score of 90% for fear-deep-phase1 dataset 1, while DT produced an f1-score of 86% for neutral-shallow-phase1 dataset 1.

B. *Physiological signals - Multi-class classification*

Table IV shows the results of several experiments that focused on specific stages of the trial. The experiments used tsfresh features calculated from `dataset 2` only and `dataset 1 & 2` and four ML algorithms for classifying individuals as having *low*, *medium*, or *high* levels of alexithymia. The best performance was achieved by MLP, LR, and RF,

TABLE III: Mean F1 (%) scores for binary classification using physiological features. (F: Fear, J: Joy, N: Neutral, D: Deep processing, S: Shallow processing, P1: Phase1, P2: Phase2)

Instance	Dataset	MLP	LR	DT	RF
Universal	Dataset 1	83	81	69	**85**
	Dataset 2	59	83	59	81
	Dataset 1 & 2	53	62	58	65
	Concatenated top features	68	69	60	77
FDP1	Dataset 1	**90**	86	80	**94**
	Dataset 2	64	**96**	58	76
	Dataset 1 & 2	69	78	61	71
	Concatenated top features	80	87	67	77
FDP2	Dataset 1	83	**86**	77	76
	Dataset 2	67	70	58	72
	Dataset 1 & 2	73	72	63	68
	Concatenated top features	72	69	62	75
JDP1	Dataset 1	86	**88**	79	85
	Dataset 2	70	84	53	51
	Dataset 1 & 2	72	68	55	74
	Concatenated top features	80	80	6	77
JDP2	Dataset 1	82	84	79	81
	Dataset 2	70	**96**	38	8
	Dataset 1 & 2	62	69	62	69
	Concatenated top features	74	70	63	77
NDP1	Dataset 1	80	79	78	75
	Dataset 2	38	**83**	56	61
	Dataset 1 & 2	62	61	55	66
	Concatenated top features	62	66	56	68
NDP2	Dataset 1	**83**	77	66	81
	Dataset 2	56	82	50	79
	Dataset 1 & 2	55	60	53	66
	Concatenated top features	71	71	57	69
FSP1	Dataset 1	**88**	**88**	80	83
	Dataset 2	76	75	70	82
	Dataset 1 & 2	72	75	63	69
	Concatenated top features	70	78	61	77
FSP2	Dataset 1	75	75	68	8
	Dataset 2	71	**96**	64	58
	Dataset 1 & 2	77	75	58	73
	Concatenated top features	65	81	62	76
JSP1	Dataset 1	**86**	**86**	74	78
	Dataset 2	58	74	48	63
	Dataset 1 & 2	66	72	56	69
	Concatenated top features	75	74	67	71
JSP2	Dataset 1	**87**	86	83	83
	Dataset 2	50	76	39	62
	Dataset 1 & 2	68	73	59	63
	Concatenated top features	69	74	55	70
NSP1	Dataset 1	82	88	86	**89**
	Dataset 2	64	63	64	73
	Dataset 1 & 2	75	68	68	75
	Concatenated top features	79	84	70	75
NSP2	Dataset 1	77	**81**	79	79
	Dataset 2	51	67	54	69
	Dataset 1 & 2	59	69	55	62
	Concatenated top features	65	67	65	69

which achieved f1-scores of 85%, 83%, and 79%, respectively. The DT algorithm had lower f1-score, achieving only 62% for the joy-deep-phase2 dataset 2.

C. *Physiological and audio signals - Multi-class classification*

We employed different approaches to process the audio signals. First, we used the original audio signals as input to pre-trained models such as wav2vec and Hubert, achieving f1-score of 42% and 38% respectively. Second, we extracted multiple audio features using the librosa library. We then combined these audio features with the tsfresh features from

TABLE IV: Mean F1 (%) scores for multi-class classification using physiological features.

Instance	Dataset	MLP	LR	DT	RF
Universal	Dataset 2	**85**	83	59	79
	Concatenated top features	75	72	45	77
FDP1	Dataset 2	77	79	56	69
	Concatenated top features	**81**	**81**	54	75
FDP2	Dataset 2	**74**	67	56	69
	Concatenated top features	73	72	54	65
JDP1	Dataset 2	70	69	51	71
	Concatenated top features	**78**	69	45	75
JDP2	Dataset 2	75	**76**	62	75
	Concatenated top features	73	64	46	66
NDP1	Dataset 2	77	75	44	**78**
	Concatenated top features	70	73	49	66
NDP2	Dataset 2	69	67	62	**73**
	Concatenated top features	55	65	42	67
FSP1	Dataset 2	**82**	67	52	74
	Concatenated top features	66	74	46	68
FSP2	Dataset 2	74	75	59	77
	Concatenated top features	75	**79**	52	74
JSP1	Dataset 2	75	**77**	62	75
	Concatenated top features	**77**	74	45	62
JSP2	Dataset 2	64	71	53	70
	Concatenated top features	**73**	68	48	63
NSP1	Dataset 2	74	75	59	76
	Concatenated top features	76	75	5	72
NSP2	Dataset 2	**84**	8	59	76
	Concatenated top features	77	68	55	59

the physiological signals. Our goal was to perform multi-class classification to represent a distinct alexithymic category. Interestingly, our results showed that the traditional ML algorithms outperformed both wav2vec and Hubert. Moreover, the merged tsfresh and librosa features performed the best among all the situations tested. This may be because the merged dataset provides complementary information that improves the performance of the alexithymia classification models.

TABLE V: Mean F1 (%) scores for physiological and audio features for the Universal model.

Datasets	MLP	LR	DT	RF
Audio features	**63**	41	49	46
Physio features	**85**	83	59	79
Audio & Physio features	**86**	85	70	75

VI. Discussion

This study aimed to explore the performance of various models and datasets for binary and multi-class classification of alexithymia. For binary classification, we tested four models: LR, MLP, DR, and RF. We found that LR achieved the best results, and generally, dataset 1 performed better than dataset 2 or datasets 1 & 2. Additionally, we observed that fear and joy imagery expression were more effective in classifying alexithymia than the neutral emotion. Notably, we found that phase 1 was generally more effective than phase 2, while depth of processing did not significantly contribute to the classification. These findings have important implications for the development of more accurate models for alexithymia classification.

For multi-class classification, we used the same models as in binary classification and found that the universal model achieved the best results for MLP, LR, and RF. We also found that dataset 2 performed better than dataset 1 or datasets 1 & 2. We observed that fear emotion was more descriptive than the other two emotions, and similarly to binary classification, depth of processing and phases were not very descriptive.

Furthermore, we investigated audio classification and found that pre-trained models and audio mfcc features were not useful in this case. However, when we combined audio and physiological features, we achieved better outcomes for the universal model compared to individual modality results.

Overall, our results suggest that LR and the universal model are effective models for binary and multi-class classification of emotions, respectively. Moreover, fear emotion is more descriptive than other emotions, and depth of processing and phases do not contribute much to the classification. Lastly, combining audio and physiological features can enhance the classification outcome.

In summary, this study has contributed to the field of alexithymia classification by providing a new approach using ML techniques. Our findings can be useful for developing better diagnostic tools for alexithymia, which can help clinicians and researchers to better understand this disorder. However, further studies are needed to validate our findings and explore other potential emotions and features for alexithymia classification. Virtual reality simulations can also be used to induce emotion processing imagery in participants, thereby generating more diverse and controlled datasets.

VII. Conclusion

This study provides novel insights into the classification of alexithymia using ML techniques, in particular by providing ML approaches for classifying alexithymia using physiological and audio signals. Our findings suggest that fear might be the most representative emotion for alexithymia classification. Furthermore, we observed that traditional ML algorithms outperform pre-trained models, indicating that models with fewer degrees of freedom might generalise better when the data is scarce. The quality of the dataset was also found to be critical in achieving good results, emphasising the need for more algorithms to further improve the classification performance. Lastly, our results indicate that multi-modality outperforms unimodal approaches, which is in agreement with previous work in this area [12]. Finally, we hope that the collected datasets will further encourage research in this direction.

Ethical Impact Statement

Ethical considerations are essential in the field of affective computing, particularly when dealing with human emotions and sensitive data. In this study, we obtained informed consent from all participants and ensured the anonymity and confidentiality of their data. We encourage researchers and practitioners in this field to continually evaluate and address ethical concerns and implications in their work, particularly regarding the privacy and well-being of individuals involved.

REFERENCES

[1] P. E. Sifneos, *Short-term psychotherapy and emotional crisis*. Harvard University Press, 1972.

[2] Z. K. Nekouei, H. T. N. Doost, A. Yousefy, G. Manshaee, and M. Sadeghei, "The relationship of alexithymia with anxiety-depression-stress, quality of life, and social support in coronary heart disease (a psychological model)," *J. Educ. Health Promot.*, vol. 3, p. 68, Jun. 2014.

[3] M. K. Yontem and K. Adem, "Prediction of the level of alexithymia through machine learning methods applied to automatic thoughts," *Psikiyatr. Guncel Yaklasimlar - Curr. Approaches Psychiatry*, vol. 11, no. S1, pp. 64–79, Jan. 2019.

[4] K. S. Goerlich, "The multifaceted nature of alexithymia – a neuroscientific perspective," *Frontiers in Psychology*, vol. 9, 2018.

[5] D. Greene, P. Hasking, M. Boyes, and D. Preece, "Measurement invariance of two measures of alexithymia in students who do and who do not engage in non-suicidal self-injury and risky drinking," *Journal of Psychopathology and Behavioral Assessment*, vol. 42, no. 4, pp. 808–825, Dec 2020.

[6] G. Panayiotou, M. Panteli, and E. Vlemincx, *Processing Emotions in Alexithymia: A Systematic Review of Physiological Markers*. Cambridge University Press, 2018, p. 291–320.

[7] G. Panayiotou, *Alexithymia as a Core Trait in Psychosomatic and Other Psychological Disorders*. Cham: Springer International Publishing, 2018, pp. 89–106.

[8] N. Farhoumandi, S. Mollaey, S. Heysieattalab, M. Zarean, and R. Eyvazpour, "Facial emotion recognition predicts alexithymia using machine learning," *Computational Intelligence and Neuroscience*, vol. 2021, p. 2053795, Sep 2021.

[9] B. Myroniv, C.-W. Wu, Y. Ren, A. Christian, E. Bajo, and Y.-c. Tseng, "Analyzing user emotions via physiology signals," *Data Science and Pattern Recognition*, vol. 2, 12 2017.

[10] T. Christy, L. I. Kuncheva, and K. W. Williams, "Selection of physiological input modalities for emotion recognition," *UK: Bangor University*, 2012.

[11] E. Constantinou, G. Panayiotou, and M. Theodorou, "Emotion processing deficits in alexithymia and response to a depth of processing intervention," *Biol. Psychol.*, vol. 103, pp. 212–222, Dec. 2014.

[12] V. Filippou, N. Theodosiou, M. Nicolaou, E. Constantinou, G. Panayiotou, and M. Theodorou, "A wavelet-based approach for multimodal prediction of alexithymia from physiological signals," in *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ser. ICMI '22 Companion. New York, NY, USA: Association for Computing Machinery, 2022, p. 177–184.

[13] G. Panayiotou and E. Constantinou, "Emotion dysregulation in alexithymia: Startle reactivity to fearful affective imagery and its relation to heart rate variability," *Psychophysiology*, vol. 54, no. 9, pp. 1323–1334, Sep. 2017.

[14] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," ser. AVEC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 3–12.

[15] ——, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," 2019.

[16] V. Kolodyazhniy, S. D. Kreibig, J. J. Gross, W. T. Roth, and F. H. Wilhelm, "An affective computing approach to physiological emotion specificity: toward subject-independent and stimulus-independent classification of film-induced emotions," *Psychophysiology*, vol. 48, no. 7, pp. 908–922, Jul. 2011.

[17] C. He, Y.-j. Yao, and X.-s. Ye, "An emotion recognition system based on physiological signals obtained by wearable sensors," in *Wearable Sensors and Robots*, C. Yang, G. S. Virk, and H. Yang, Eds. Singapore: Springer Singapore, 2017, pp. 15–25.

[18] S. Salari, A. Ansarian, and H. Atrianfar, "Robust emotion classification using neural network models," in *2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, 2018, pp. 190–194.

[19] P. J. Bota, C. Wang, A. L. N. Fred, and H. Plácido Da Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140 990–141 020, 2019.

[20] A. Vulpe-Grigoraşi and O. Grigore, "A neural network approach for anxiety detection based on ecg," in *2021 International Conference on e-Health and Bioengineering (EHB)*, 2021, pp. 1–4.

[21] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," 10 2017, pp. 811–819.

[22] Y. Tan, Q. Zeng, and H. Zhang, "Research on anxiety detection based on personalized data markers," *Journal of Physics: Conference Series*, vol. 1948, no. 1, p. 012035, jun 2021.

[23] F. Demir, N. Sobahi, S. Siuly, and A. Sengur, "Exploring deep learning features for automatic classification of human emotion using eeg rhythms," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 14 923–14 930, 2021.

[24] R. Elalamy, M. Fanourakis, and G. Chanel, "Multi-modal emotion recognition using recurrence plots and transfer learning on physiological signals," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2021, pp. 1–7.

[25] C. V. Witvliet and S. R. Vrana, "Psychophysiological responses as indices of affective dimensions," *Psychophysiology*, vol. 32, no. 5, pp. 436–443, Sep. 1995.

[26] C.-C. Hsiao, "Polyphase filter matrix for rational sampling rate conversions," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, 1987, pp. 2173–2176.

[27] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[28] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[29] G. Strang, "Wavelets," *American Scientist*, vol. 82, no. 3, pp. 250–255, 1994.

[30] E. O. Brigham and R. E. Morrow, "The fast fourier transform," *IEEE Spectrum*, vol. 4, no. 12, pp. 63–70, 1967.

[31] N. Ricker, "The form and laws of propagation of seismic wavelets," in *3rd World Petroleum Congress*. OnePetro, 1951.

[32] A. Gholamy and V. Kreinovich, "Why ricker wavelets are successful in processing seismic data: Towards a theoretical explanation," in *2014 IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES)*. IEEE, 2014, pp. 11–16.

[33] P. E. McKnight and J. Najab, "Mann-whitney u test," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

[34] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[35] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub, and C. Cléder, "Automatic Speech Emotion Recognition Using Machine Learning," in *Social Media and Machine Learning [Working Title]*. IntechOpen, Mar. 2019.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[40] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 2019.

[41] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.