Can Machine Learning Models Recognise Emotions Better Than Humans?

Student Name: Jeffin Siby
Supervisor Name: Prof. Effie Lai-Chong Law
Submitted as part of the degree of BSc Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—Audio and video are popular modalities used in the domain of emotion recognition. Machine learning (ML) techniques such as SVMs and artificial deep neural networks have shown impressive results in inferring human emotion. This project compared the performance of a state-of-the-art (SOTA) ML model for each of the two modalities to the performance of humans. We also explored the items most commonly classified as 'neutral', and compared the results with that of humans. Using the CREMA-D audio-visual emotion dataset, a CNN-LSTM model and SVM model were trained for both video-only (visual) and audio-only time-series data respectively. Their performance was evaluated both on a matching test set (i.e., stimuli for which the binomial majority of human rater's emotion classification votes matched that intended by the actor) and also on a test set that contained all non-matching (i.e. items for which the binomial majority of human rater's emotion classification vote did not match the emotion intended by the actor and those where there was no binomial majority). Accuracy, F1-score, precision and recall were used on a macro-averaged and per-class basis to evaluate both human and ML results. For all the macro-averaged metrics, the ML models outperformed humans for both audio and visual data. The difference was more evident for the non-matching test set. The CNN-LSTM model achieved an accuracy of 80.8% and 39.0% on the matching and non-matching visual data respectively which was higher than the human accuracy of 75.9% and 19.4%. Similarly, the audio classification results showed 81.0% and 34.3% accuracy for the SVM model on matching and non-matching test sets, whereas human accuracy was 68.9% and 17.9%. The performance of ML models on the test data highlights their advantage over humans in the domain of monomodal emotion recognition with audio and visual data, particularly for emotions that humans have difficulty classifying.

Index Terms—Facial emotion recognition, Speech emotion recognition, CREMA-D, monomodal, CNN-LSTM, SVM

1 Introduction

Affective computing is an interdisciplinary field of research that is concerned with enabling intelligent systems to sense a human's emotional state [1]. This field originated with Rosalind Picard's 1995 paper [2] on affective computing. Initial attempts of unimodal emotion recognition, predominantly via facial expressions, have been carried out in the past [3], [4], which has now extended to explore other modalities such as audio, gesture and psychological signals.

Defining the number of existing emotions and categorising them is an ongoing debate among researchers. A common set of categories for emotion that is seen in the literature are 'happiness', 'surprise', 'anger', 'disgust', 'sadness' and 'fear' which psychologist Ekman [5] proposed. Ekman believed these to be independent of factors such as gender, race, education and ethnicity [6], making them globally recognisable.

However, some disagree with the 6 categories and instead suggest upwards of 20 multimodal expressions [7]. Others disregard the discrete categories altogether and suggest a continuous scale consisting of affective features such as valence and arousal [8].

In this paper, we will assume the traditional 6 categories as they have been the primary focus of research for decades and thus has a vast range of related scientific papers which can be used for analysis [9].

The most popular modalities used to infer emotion are audio and visual data. For visual data, a common metric to label and identify certain muscle contractions (often for the purpose of emotion inference), such as lowering or raising the inner corners of the eyebrows, are Action Units (AUs) which are defined the Facial Action Coding System (FACS). These values can be combined or used individually to anatomically identify facial expressions and help aid the inference of an emotion. Although AUs are analysed independently, it is also important to note that many facial movements are anatomically wired together such that certain muscle contractions require the movement of other muscles, thus there can be a dependency between certain AUs [9].

Audio features could also be used to help infer human emotion. Studies [10], [11], [12] have reported that pitch may be the most essential feature in the context of emotion communication, followed by other features such as duration and loudness. Furthermore, one study [9] on children suggested that expressions of anger were more evident from vocal observations than facial. It was seen that children raised their voices 42% of the time when faced with a situation evoking anger, in contrast to the scowling facial configurations which they made only 16.2% of the time [9]. This could suggest that vocal cues may have significant involvement in expressing an emotion – potentially more significant depending on the emotion (e.g. anger) and age. Nevertheless, it's also important to note that the relationship between audio features and emotions may vary as individuals age due to emotion regulation. For example, it is

seen to be less socially acceptable for people to raise their voices when they're angry, thus people may regulate their emotions to be more subtle as they grow older.

1.1 Challenges with inferring emotion

As with many areas in the field of affective computing, inferring emotion comes with a lot of potential issues that we must be aware of when researching the analysis of emotional expression and perception.

Specificity: In fact, it should be questioned whether it is even possible to directly infer human emotion from a set of audio-visual (audio and visual) data. For example, Although Beaupré et al. [13] and Ekman [5] suggest basic emotions are universally recognisable based on specific expressions, Keltner and Cordaro [14] claim that 'there is no one-to-one correspondence between a specific set of muscular facial actions or vocal cues and any and every experience of emotion'.

Contextualisation: There is also a lot of debate on the extent of variation in attributes that indicate a particular emotion. Some propose the 'basic-emotion approach' framework which suggests facial movements vary to a small degree around a set of core movements [9]. Other frameworks suggest a substantial variation depending on factors such as context. For example, a study [15] observed that people who won medals only smiled when facing an audience, e.g. when standing on a podium, as opposed to facing away from them. Other internal factors such as a person's metabolic state and past experiences that are present when an emotion is invoked can also have an effect as facial muscles are tied to the immediate context [9].

Enculturation: Many instances of literature also suggest potential variations in emotion expression across various cultures. For example, as reported by Eibl-Eibesfeldt [16], a rapid eyebrow raise may be observed in the US and Europe when greeting friends, whereas, in Japan, it serves as an impolite method of greeting. However, it is also possible that a subset of emotions are recognisable across cultures. For example, a study by Cordaro et al. [17] asked college students from China, India, Japan, Korea and the US to pose facial expressions given a scenario and the similarity between FACS-coded expressions were calculated. The study found strong to moderate evidence that participants across cultures share common beliefs about the expressive pose for the anger, fear, and surprise category, but weak evidence of common belief for the disgust and sadness categories. This could suggest that some emotions may have better recognition across cultures than others and may be easier to predict. Hence, when collecting and using data, it is critical that we have data from a diverse set of people to eliminate bias and also to prevent getting an unrepresentative metric for accuracy.

Methodological challenges: Issues with data collection techniques should also be considered when selecting datasets. Crowdsourcing is a popular option when presented with the task of labelling data as it enables the recruitment of a large number of raters. This can mitigate individual differences of raters and thus individual rater biases. However, it has risks of being subjective to a particular demographic if extra care is not put in to ensure that the raters come from diverse backgrounds. We must also ensure that the rating quality of the raters is to a high standard, possibly though reliability metrics such as 'Krippendorff's alpha' (used by e.g. Cao et al. [18]).

Another option is to ask participants to self-report. However, this technique is frowned upon as its not only highly subjective to the participant's judgement about their own feelings, but people can also be susceptible to experiencing an emotion without being fully aware of it or even have difficulty expressing how they felt in words [9].

These challenges have been taken into consideration when selecting an appropriate dataset to train and test the ML models.

1.2 CREMA-D

The dataset chosen for this project was the CREMA-D audio-visual emotion dataset as it provides individual human rater classification results for each audio-only and video-only stimuli, which can be used to compare with ML models. The dataset contains 7,442 clips of 91 actors who were asked to read out 12 sentences representing the emotions 'happy', 'anger', 'neutral', 'fear', 'disgust' and 'sad' (similar to Ekman's six [5], but with the exclusion of 'surprise' and inclusion of 'neutral') [18]. These clips were approved by directors and judged by raters via crowdsourcing [18]. Each clip is annotated with the intended emotion class and the individual emotion labels given by raters, as well as the average intensity (continuous scale) of the clip given by the raters [18]. For the 'neutral' clips, the confidence level was given on a scale of 0-100 instead of intensity.

As explained by Cao et al. [18], the following three subsets are included in the final dataset:

- Matching: where the group perceived emotion matches the intended emotion.
- Non-matching: where the group perceived emotion differs from the intended emotion.
- Ambiguous: where the group perceived emotion doesn't have a majority.

Out of the 7442 stimuli for each modality, the matching, non-matching and ambiguous audio-only subsets contained 41%, 46% and 13% of the responses respectively [18]. The video-only subsets contained more responses in the matching subset, with 64%, 25% and 11% of responses in the order of subsets stated previously [18].

For our study, the non-matching and ambiguous subsets have been combined into a single non-matching set for audio and visual data independently. This set represents items which humans found particularly difficult to infer emotion from. The distribution of instances for each emotion class can be seen in Appendix Fig. 16 and Fig. 17 (note that the non-matching audio set contained more instances than the matching set due to the large number of misclassifications).

The primary motivation for this study is the low accuracy of humans on recognising emotion from the audio and visual stimuli. It was reported that human raters attained an accuracy of 49% on the entire audio data, 58.2% on all the video-only data and 63.6% accuracy when shown stimuli that contained both audio and video [18]. The

surprisingly low accuracy for the two monomodal stimuli raises the question of whether ML models can perform better at classifying these emotions.

1.2.1 Misclassification of 'neutral'

Another interesting phenomenon for the CREMA-D dataset that has not been addressed by others performing ML experiments is the large percentage of misclassifications of the 'neutral' emotion. For example, when looking at the entire set of audio data, it was seen that 48% of the ratings for the 'sad' stimuli and 45% of the ratings for the 'happy' stimuli were misclassified as 'neutral'.

This ambiguity for the term 'neutral' is prevalent throughout the literature with different opinions on what it actually is and is another motivation for this study. Although we do not discuss what 'neutral' is and whether it is an emotion category, we explored how ML models compare to humans and if they make similar misclassifications of the 'neutral' class as seen for humans. We also explored which emotions are most commonly misclassified as 'neutral' to help shed more light on the open debate.

1.3 Project objectives

This project will use SOTA emotion classification techniques and build two SOTA ML model classifiers and compare their performance against human raters. Accuracy, F1-score, precision, and recall will be used to compare the collective human performance metrics with that of the ML models for each of the time-series visual and audio testing data. This project answers the following key sub-questions to answer the research question 'Can Machine Learning Models Recognise Emotions Better Than Humans?':

- How does the performance of ML models compare to human raters for audio data?
- How does the performance of ML models compare to human raters for visual data?
- Is the accuracy of ML models lower on non-matching items as humans?
- Do ML models tend to misclassify a large percentage of items as 'neutral' as humans?

The novelty of this project is twofold. Firstly, the separation of data into two sets (matching and non-matching) to compare ML performance to human performance on items is not in any other paper. This separation allows the model to be evaluated on a more challenging test set so that the model performance can be compared to data where humans had a high accuracy in classifying and also to data which humans found particularly difficult to classify. Secondly, no other paper using the CREMA-D dataset provides a per-emotion analysis comparing ML models to human classification results whilst also investigating the trends in emotions being misclassified as 'neutral'.

It is important to note that due to major differences between the audio and visual data processing including the algorithms used and the amount of data used for training and testing, it was not appropriate to compare the audio ML model and video ML model and say one performs better than the other purely off metrics. Therefore, the development and analysis of both models were kept independent of each other throughout this project.

Moreover, although the CREMA-D dataset provides multimodal audio-visual data, this research project did not pursue the multimodal scenario due to two primary reasons. Firstly, constructing a SOTA fusion model and optimising it along with the monomodal models was deemed too demanding given the time constraints imposed by the project. Secondly, the monomodal scenarios were selected over the multimodal case because, as detailed in the 'CREMA-D' subsection, the monomodal stimuli produced the lowest accuracies for human performance, hence highlighting the potential for improvement through the assistance of ML models.

To aid future researchers, this paper also provides the results of various experiments carried out to improve the implementations of the SOTA ML techniques. This includes the use of principal component analysis (PCA) for reducing the dimensionality of the audio data, which not only improved the accuracy of the SVM model but vastly improved computational complexity.

1.4 Assumptions

One of the biggest assumptions in this study is that the actors in the CREMA-D dataset portray the ground truth for the emotion intended in the clip. We also assume that these emotions are generalisable to the public.

Additionally, we also assume that the raters involved in the CREMA-D study are representative of the general public as the collective performance of these raters across each test set was used as 'human' data to compare with the ML model results.

2 Related Work

This section presents an overview of the chosen dataset and the results of the search for a second test dataset. The survey on related emotion recognition techniques used in the literature is also provided and is split into two sections to discuss SOTA visual classification and audio classification techniques independently.

2.1 Datatsets

Although several human emotion datasets exist, the search for an appropriate dataset has been limited to datasets containing audio and visual data as these are the most popular modalities for inferring emotion. This project will only focus on categorical emotion classification data as this is what's used in the CREMA-D dataset.

In the domain of emotion recognition, there are 3 main categories into which datasets fall into [19]:

- Acted (Simulated)
- Elicited (Induced)
- Natural

'Acted' datasets record actors (professional or semi-professional) who perform or provide an utterance of a specific emotion in a studio environment. The CREMA-D dataset is an example of an 'acted' dataset.

'Natural' datasets on the other hand are obtained from a 'natural' setting in which spontaneous emotions are captured. This can include recordings from talk shows, call-centre recordings [19], YouTube and similar sources.

'Elicted' datasets fall in-between both and are created by placing participants that are to be recorded in a simulated emotional situation [19].

Information about the diversity of participants was also provided by Cao et al. [18], where they stated that 73.60% of raters and 58.24% of the actors were 'Caucasian' in comparison to other ethnicities such as 'Asian', which contributed to only 4.50% of the raters and 7.69% of the actors. With the previous discussion regarding variation in emotional expression across cultures in mind, a potential limitation of this dataset could be that a model trained on the data may not generalise well when tested on a non-western audience.

To test the validity of the ML model results on other data, a search for datasets similar to CREMA-D (i.e. with the same set of categorical emotion labels) was carried out. A summary of the explored audio-visual emotion datasets is provided in Table 1.

The major reason that other datasets were not chosen for this study is primarily due to the non-existence of human results, without which a comparison to ML models cannot be drawn. The CREMA-D dataset on the other hand provides a detailed breakdown of the results of individual raters for each stimulus, making it possible to compare to ML performance on said stimuli. Additionally, many datasets collected data from raters who responded to multimodal stimuli instead of considering each modality individually. Some datasets, including CMU-MOSI [20], SEND [21], RECOLA [22], SEWA [23], VAM [24] and HUMAINE [25] used a continuous valence/arousal scale as opposed to a categorical classification used in CREMA-D. For datasets that used categorical, there were several issues that made them incompatible for comparison, for example, IEMOCAP [26] used 8 states but did not contain the emotion of 'disgust'. Others such as the SEWA dataset contained undesirable data including commercial videos without any humans in them. The process of removing such data would require significant manual effort, thus the exploration for a second test dataset was discontinued.

2.2 Visual Classification Techniques

Conventional approaches for Facial Emotion Recognition (FER) follow the 3 major steps [38]:

1) Face detection and landmark detection on input images or frames.
2) Feature extraction.
3) Facial emotion classification.

Ghimire and Lee [39] took a conventional approach and calculated the Euclidean distance and angle of 52 facial landmarks before either using multi-class AdaBoost with dynamic time warping or SVM on the boosted feature vectors [38].

Another popular method is to use a FACS [40] or alternatively a histogram of oriented gradients (HOG) feature map, common in the field of computer vision for object detection, before again using classifiers such as SVM, AdaBoost and random forest [38]. Chen et al. [41] proposed such a FER system by first performing face detection and then extracting the brows, eyes, nose and mouth from the face region. They then extracted HOG features since they are sensitive to object deformations [41] (facial muscular contractions in this case) by capturing information about the local gradient intensity and orientation of image subsections, and then proceeded to train and test an SVM model using this feature set. The proposed facial recognition pipeline is shown in Fig. 1.

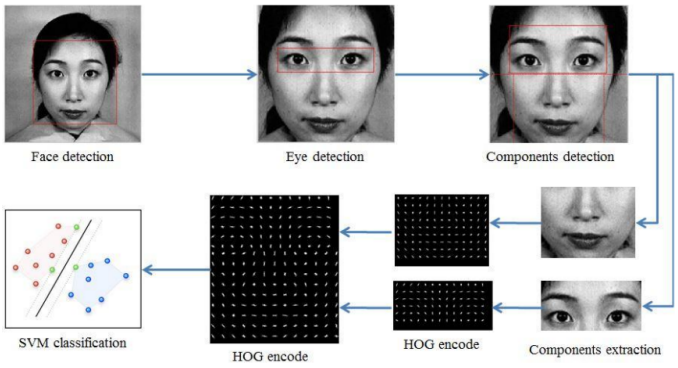

Fig. 1. Proposed facial recognition system using HOG feature descriptors by Chen et al. [41].

However, deep learning techniques are now most commonly used in literature [42], yielding SOTA results. CNN architectures are the most popular as they perform end-to-end learning directly from the input images [38].

For example, Ristea et al. [43] trained a 7-layer CNN model consisting of convolution layers, pooling layers, and a fully connected layer, with 2 frames per image as input and achieved 62.84% accuracy on the CREMA-D dataset.

Nevertheless, although the performance of CNN architectures are remarkable, its inability to reflect temporal variations in time-series data such as videos have led to hybrid approaches that combine CNN architectures with a memory component such as RNNs or LSTMs [38]. This hybrid approach allows both spatial and temporal information to be captured.

Recent work by Ryumina et al. [40] explored the performance of such CNN hybrid models on various emotion datasets that cater for FER. The analysis of results concluded that CNN-LSTM and CNN-LSTM-A (a variation of the CNN-LSTM model with the main change being the addition of the attention mechanism proposed by Winata et al. [44]) provided the highest accuracies compared to other hybrids including CNN-GRU (which has Gated Recurrent Unit (GRU) layers instead of LSTM layers) and CNN-SVM on the RAMAS [45], RAVDESS, CREMA-D and AffWild2 [46] datasets. The average uniform average recall (UAR) of the CNN-LSTM, CNN-LSTM-A, CNN-GRU and CNN-SVM models across all evaluated datasets were 53.9%, 53.6%, 52.4% and 40.2% respectively. The low-performance scores demonstrate that even advanced ML models have difficulty classifying emotion.

Hans and Rao [47] also achieved SOTA results on the CREMA-D dataset by implementing a CNN-LSTM architecture. They trained proposed CNN-LSTM models with 5, 6 and 7 layers with an input of 75 frames per video. The highest test accuracy of 78.5% was achieved by a 6-layer model. Due to the outstanding results, their proposed method was used for the visual classification approach in this paper.

TABLE 1
Summary of results audio-visual emotion datasets. # States means number of states, Cont. means that the dataset used continuous values such as valence/arousal/dominance.

Dataset	# States	Language	Modalities	Type
CMU-MOSI	Cont.	English	Language/Audio/Visual	Natural
SEND	Cont.	English	Audio/Visual	Elicited
RECOLA	Cont.	French	Audio/Visual	Natural
EmoReact [27]	17	English	Audio/Visual	Natural
GEMEP [28]	18	French	Audio/Visual	Acted
SEWA	Cont.	English/Chinese/German/Greek/Hungarian/Serbian	Audio/Visual	Natural
VAM	Cont.	German	Audio/Visual	Natural
HUMAINE	Cont.	English/French/Hebrew/Other	Audio/Visual/Other	Acted/Elicited/Natural
IEMOCAP	8	English	Audio/Visual	Acted
OMG [29]	6	English	Audio/Visual	Natural
RAVDESS [30]	8	English	Audio/Visual	Acted
CMU-MOSEI [31]	6	English	Language/Audio/Visual	Natural
SAVEE [32]	8	English	Audio/Visual	Acted
CHEAVD [33]	8	Mandarin	Audio/Visual	Acted/Natural
eNTERFACE'05 [34]	6	English	Audio/Visual	Elicited
SEMAINE [35]	Cont.	English/Greek/Hebrew	Audio/Visual	Natural
AFEW [36]	7	English	Audio/Visual	Natural
MOUD [37]	3	Spanish	Language/Audio/Visual	Natural

2.3 Audio Classification Techniques

Trends in audio classification techniques also tend to follow one of two approaches. The first is to use dedicated feature extraction software such as COVAREP [48], OpenEAR [49] or openSMILE [50] to extract hand-engineered features such as pitch, cepstral coefficients, zero-crossings etc. [51] which are related to emotion and the tone of speech. The extracted features are then used to train ML models such as SVMs or logistic regression models after any necessary pre-processing.

With the rise in popularity of neural networks, the second group of techniques take advantage of visual representations of audio such as mel-spectrograms and use neural network architectures such as a CNN to perform automated feature extraction and classification. Often, mel-spectrograms are created based on the application of the fast Fourier transform (FFT) on the input audio data, which results in a time-frequency representation [52]. These spectrograms are better suited for applications that need to model human hearing perception, thus they are widely used in the domain of speech emotion recognition (SER).

CNNs have shown promising results in various SER experiments. For example, the CNN model proposed by Wani et al. [52] achieved an accuracy of 79.4% on the SAVEE dataset and achieved an even higher accuracy of 87.8% when using a novel Deep Stride CNN (DSCNN) which uses different kernel stride values for downsampling instead of using pooling layers. Badshah et al. [51] also used spectrograms as input to a CNN classifier (see Fig. 2), achieving an accuracy of 84.3% on EMO-DB [53].

It should be noted that some studies, such as that conducted by Gokilavani et al. [54] have suggested that variants such as CNN-LSTM models could perform worse than CNN models in the domain of SER. Fayek et al. [55] have also conducted investigations to test the performance hybrids including various LSTM-RNN models and came to the conclusion that the CNN architectures achieved better performance.

Although deep learning techniques are on the rise, SVMs trained with specialised features tend to outperform most deep neural networks such as RNNs and (multi-layer perceptrons) MLPs in the domain of SER using audio data and is one of the most used classifiers in the literature [56]. For example, Özseven [56] demonstrated that an SVM model trained on a particular feature set obtained accuracies of 84.62%, 60.40%, and 72.39% on the EMO-DB, EMOVO [57] and SAVEE datasets respectively. Comparatively, a MLP classifier trained and tested on the same datasets using the same feature set resulted in lower accuracies of 81.32%, 58.58% and 71.17% respectively.

Further, experiments on the CREMA-D dataset itself have added supporting results to the superior performance of SVMs. Singh et al. [58] demonstrate how an SVM classifier outperforms an 6-layer RNN classifier. The authors claim that SVM has an overall higher performance score to the RNN classifier, particularly for the emotions of 'Anger', 'Fear', and 'Sad' where the SVM model achieved accuracy scores of 90.13%, 84.35% and 85.43% compared to the RNN model which achieved lower accuracies of 87.79%, 83.43% and 83.43% respectively. It should also be noted that the average time taken for the SVM classifier to run on the test data was approximately 15.75s. In contrast, the RNN model took an average of 1594s to run [58], which exhibits the computational complexity of neural networks compared to traditional models such as SVM classifiers.

Valles and Matin [59] validated the superior performance of SVMs on the CREMA-D dataset using 62 features generated from a custom feature set, and showed that the SVM model achieved an accuracy of 57.6% on clean audio data from the CREMA-D dataset, followed by an MLP model which achieved 54.7% accuracy and finally an RNN model which achieved 55.0% accuracy. This order of model performance remained consistent when also tested on the clean RAVDESS dataset, resulting in an accuracy score of 80.6% for SVM, 77.6% for MLP and 61.2% for RNN.

3 Methodology

The solution developed to answer the research question focuses on the task of emotion classification through audio-only and video-only input data independently. By compar-

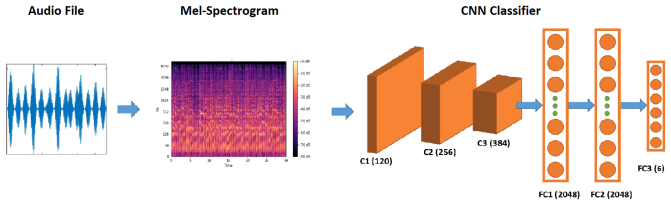

Fig. 2. CNN classifier pipeline with mel-spectrogram adapted from Badshah et al. [51].

ing the results of SOTA ML models against humans, this project aims to answer the research question by exploring if better performance results can be achieved by the models. The SOTA techniques used for both classification pipelines are explained in this section, including the results of experimentation, and details on techniques to improve the accuracy of the models.

3.1 Train-Test split

For both audio and visual data, the provided dataset was split into 'matching' and 'non-matching' sets. This was to allow an evaluation to be conducted to compare the performance of the chosen ML models on data that humans have difficulty classifying. If the ML models performed well on the 'non-matching' test set, this would give more support in favour of the claim that ML models can recognise emotions better than humans, particularly in conditions where there is a lot of human ambiguity.

Once this was completed, the 'matching' audio data underwent an 80/20 train/test split. This split was not only chosen because it's a popular split found in the literature, but it also ensures that there is a sufficient number of minority class instances in the test set. The 'non-matching' set was left untouched and was used as a second test set. The 80% training data was later used to train the model using a stratified 10-fold cross-validation technique in order to test the model's generalisability given the limited data (see section 3.2.2 for more details). This split resulted in 2479, 620 and 4343 audio stimuli for the matching training set, matching test set, and non-matching test set respectively.

The visual 'matching' data underwent a 70/15/15 train/validation/test split. The size of the training set was set to 70% unlike the audio data because the 'matching' visual test set contained more data entries than the audio matching set. The 70% split allowed a similar amount of training data to be used for the visual model whilst still keeping the testing data reasonable in size. A validation set was used for the visual data because, unlike the audio data, k-fold cross-validation was not performed. This is primarily due to the excess computational cost that would result from the model used. Although k-fold cross-validation was considered for the visual data, the intended neural network architecture chosen for this project would need to be trained for several days to get a suitable result. This paired with a grid-search to select the best hyperparameters would result in the network being trained for weeks. Due to the 24-hour session limit of the NCC cluster available through Durham University, this approach was not deemed feasible and so a train/validation/test split was chosen.

The split resulted in 3451 video-only stimuli for the training set, 740 for the matching validation set, 740 for the matching test set, and 2511 for the non-matching test set.

To keep the class distribution the same in all the splits for each modality, a stratified split was used. This was to ensure that each split contained instances of every class so that the model can be trained and evaluated on every class.

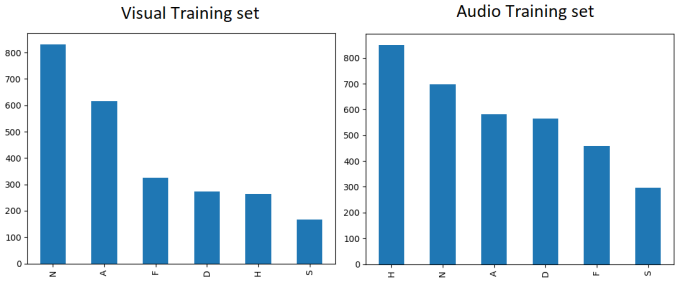

Fig. 3. Charts showing the class distributions for the audio and visual train splits before oversampling. The test sets followed the same distribution. N=Neutral, A=Anger, F=Fear, D=Disgust, H=Happy, S=Sad.

3.2 Audio classification pipeline

3.2.1 Audio Preprocessing

As stated in the section 2, SOTA results in SER are generally achieved via SVM or CNN architectures. Although many of the latest works in the literature favour the use of CNNs, an SVM architecture was chosen for this study. The SVM model was chosen due to the limited size of audio data available as a result of how the data was split into 'matching' and 'non-matching' sets. The key problem with selecting a deep learning model such as a CNN is that it requires a vast amount of training data to perform well, which was not available for the audio case. Thus the use of a CNN network was not deemed appropriate.

Although Valles and Matin [59] used a custom audio feature set consisting of 36 low-level descriptors (LLDs) to achieve good results with SVM on the CREMA-D dataset, Matin referred to the openSMILE toolkit and explained that the toolkit was not used because there were 'no Python libraries that allow users to access the features of OpenSMILE' [60] at the time of writing. However, since then, the openSMILE toolkit [50] has been made accessible via a python library and was used in this study. This toolkit is popular in the domain of SER and allows a plethora of relevant features to be extracted.

The main feature sets available through the toolkit are the ComParE_2016, GeMAPS and eGeMAPS feature sets. The INTERSPEECH Computational Paralinguistics ChallengE (ComParE) set contains 6,373 features including: 'energy, spectral, MFCC and voicing-related LLDs as well as LLDs including logarithmic harmonic to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness' [61] (see Fig. 4 for a breakdown of the LLD categories). GeMAPS on the other hand has 62 parameters consisting of 18 LLDs (including frequency, energy and spectral parameters); as well as functionals applied to these LLDs such as the mean, range and other properties; and also 6 temporal features with their assigned functionals including the mean length of voiced regions and the rate of loudness

peaks [62]. With the proven success of cepstral parameters in the modelling of affective states, the eGeMAPS extended the GeMAPS feature set and contains an additional 7 LLDs along with functionals such as arithmetic mean and the coefficient of variation etc. applied in voiced and unvoiced regions, resulting in a total of 88 parameters [62].

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	spectral
MFCC 1–14	cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F_0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice qual.
log. HNR, Jitter (local & δ), Shimmer (local)	voice qual.

Fig. 4. Table showing the 65 LLDs contained in the ComParE_2016 acoustic feature set from Schuller et al [63]. Functionals are applied to the LLDs and change in LLDs to attain the full feature set.

As concluded by Eyben et al. [62], the classification of SVMs trained with these three feature sets provide comparable results and the best-performing feature set varies with the dataset. Therefore, all three feature sets were used for experimentation in this project and their results are reported later in section 3.3.1. Our 'original' model (as referred to in section 3) initially used the ComParE_2016 feature set as this had the largest number of relevant features.

Although feature selection methods such as ANOVA are often applied after extraction [64], this method was not deemed fit for this task because the features extracted using the openSMILE library are a collection of specialised features curated for speech analysis, and removing selective items could result in the loss of valuable information.

Once the feature set had been extracted from the raw audio .wav files, they were standardised to keep the contribution of all features equal (particularly important for SVMs as they are based on distance measurements and so are affected by the range of features). Although using scikit-learn's [65] 'MinMaxScalar' (which scales the minimum and maximum values to be 0 and 1 respectively) is a popular method, the 'StandardScalar' function was chosen because the features were mostly normally distributed [66].

Due to the uneven distribution of class labels, a Synthetic Minority Oversampling Technique (SMOTE) [67] was used, which creates a synthetic instance of the minority class by selecting a point between an example from the minority class and a randomly selected nearest neighbour of that example. SMOTE was added as a step in the Python 'imblearn' [68] modules' 'Pipeline' method to fix the imbalance instead being applied directly to the full 80% train split. This was to ensure that SMOTE wasn't used to generate synthetic instances of the validation set during the stratified k-fold cross-validation to keep it truly independent of the training data.

3.2.2 SVM Training and Hyperparamter-tuning

Following this, the stratified 10-fold cross-validation with random shuffling (to keep the splits random) was performed with a grid-search to find the optimal parameters for the SVM model. The value of 10 was chosen for the number of folds as it's prevalent in the literature and generally results in a model skill estimate with low bias and a modest variance [69]. Furthermore, a stratified cross-validation approach was selected as opposed to a regular cross-validation technique to ensure there were no training subsamples containing no instances of a class.

Hyperparameter tuning was performed by altering the C and gamma values of the SVM model which alters the way in which the decision boundary that separated classes is defined. As our training data was limited in size, it was probable for overfitting to occur, and thus, the hyperparameters were selected to minimise this possibility.

The C value is a regularisation parameter that controls the trade-off between achieving a low training error and a low testing error. A lower C value will encourage the decision function's margin to be larger, resulting in a simpler decision function [65]. A larger margin would make the classifier more lenient, allowing it to be less sensitive to outliers (i.e. more tolerant to misclassifications in the training data). This could improve the generalisation of the model, therefore reducing the effects of overfitting.

The gamma value determines to what extent the decision boundary is affected by features [65]. A high gamma results in the model being influenced more by individual features, thus causing the decision boundary to overfit to the training data. Lowering gamma can reduce training accuracy by lessening the decision boundary's ability to capture data complexity, but it can improve generalisation by decreasing feature sensitivity (i.e. less prone to overfitting).

With this information, a range of values with a focus on small values for C and gamma were selected. Grid-search was performed on the following C values: 10,100,1000, and the following gamma values: 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1 and was optimised for accuracy. For building the SVM model, we used the scikit-learn [65] implementation of SVM.

After the first round of training, it was seen that for the majority of values used, the SVM model was achieving 100% training accuracy which is a clear sign of overfitting as predicted. To address this, a second grid-search was performed with the same gamma values, but with even smaller C values: 1e-4, 1e-3, 1e-2, 1e-1, 1.

Despite this, the SVM model was still achieving 100% training accuracy. Therefore data augmentation and PCA were carried out as mentioned in the subsection below in a further attempt to reduce overfitting.

The best results using approach stated thus far (see the performance of the 'original' model in Table 2) were achieved with an SVM model that had hyperparameters of C and gamma set to 10 and 1e-4 respectively.

3.3 Experiments with audio classification

3.3.1 openSMILE Feature Sets

To explore the effects of extracting different specialised features, the model was also trained and tested with the

GeMAPS and eGeMAPS feature sets provided by the openSMILE toolkit [50], and a grid-search was performed to identify the optimal parameters.

The results (see Table 2) show that an SVM model trained on the ComParE_2016 feature set resulted in considerably higher accuracy than when trained on the other feature sets. Due to the stark difference in accuracy, all further experimentation was carried out using the ComParE_2016 feature set.

TABLE 2
Table showing the accuracies achieved by the optimised SVM models trained and tested using the ComParE_2016 ('original' model), GeMAPS and eGeMAPS feature sets on the matching and non-matching test sets.

	CV Acc.	Matching Acc.	Non-matching Acc.
original	**79.06**	**80.32**	**33.43**
GeMAPS	66.92	68.87	27.49
eGeMAPS	68.25	70.32	28.97

3.3.2 Data Augmentation

In an additional attempt to reduce the effects of overfitting, the model was trained on augmented data. Popular methods of data augmentation in the domain of SER include noise addition, random shift, random change in pitch, random change in loudness, silence removal etc. [70]. Augmentations such as pitch and loudness were avoided as these features are extracted via openSMILE [50] and are directly used to help infer emotion - randomly altering these key features could negatively impact the model's performance. Silence removal was also not used because each audio clip in the CREMA-D dataset is reduced to a single utterance, thus silence removal would have been redundant.

Instead, random shifts of the audio data in either direction (left or right) by a random duration between 0 and 0.5 seconds (this range seemed appropriate given the short duration of each clip) were performed for each audio clip, and a simple Additive White Gaussian Noise (AWGN) was added at a signal to noise ratio (SNR) of 5 and 10 as is common in the literature [70], [59].

As explained by Wijayasingha [71], the SNR can be defined as:

$$SNR = 10\log\left(\frac{RMS^2_{signal}}{RMS^2_{noise}}\right) \tag{1}$$

Which can be rearranged to find RMS_{noise}:

$$RMS_{noise} = \sqrt{\frac{\sum (n_i)^2}{n}} = \sqrt{\frac{RMS^2_{signal}}{10^{SNR/10}}} \tag{2}$$

Since the mean of AWGN is 0, the standard deviation of the noise can be written as:

$$STD_{noise} = \sqrt{\frac{\sum (n_i - \mu_{noise})^2}{n}} = \sqrt{\frac{\sum (n_i)^2}{n}} = RMS_{noise} \tag{3}$$

which is, by definition, the RMS_{noise}. Therefore, the noise was generated from a Gaussian distribution with a mean of 0 and a standard deviation defined in eq. (2), and added to each audio clip in the training set.

A custom pipeline function was created for GridSearchCV [65] to ensure that the validation sets contained the clean (i.e. raw) audio data, whilst the model is trained with the features extracted from the augmented data.

The results of the models trained on audio data augmented with a random noise shift in addition to a noise addition with SNR 5 and 10 will be referred to as shno_5 and shno_10 respectively (see Fig. 5 to see an example of how a sound wave is transformed with such augmentation).

The results (Table 3) showed that the original model trained with clean data achieved a cross-validation accuracy score of 79.06%, which was more than 5% higher than the accuracies achieved by the models trained with data augmentation. This coincides with the findings of Valles and Matin [59], and Atmaja [70] who suggest that data augmentation techniques could result in lower accuracies depending on the dataset. Since a lower cross-validation accuracy was achieved with the data augmentation technique while the training accuracy remained at 100%, the clean data was used for the final model.

TABLE 3
Table showing the accuracies achieved by the SVM model trained using no data augmentation (original), data augmented with a random shift and SNR of 5 added (shno_5), data augmented with a random shift and SNR of 10 added (shno_10), and tested on the matching and non-matching test sets.

	CV Acc.	Matching Acc.	Non-matching Acc.
original	**79.06**	**80.32**	**33.43**
shno_5	73.54	71.45	28.57
shno_10	73.66	72.06	28.89

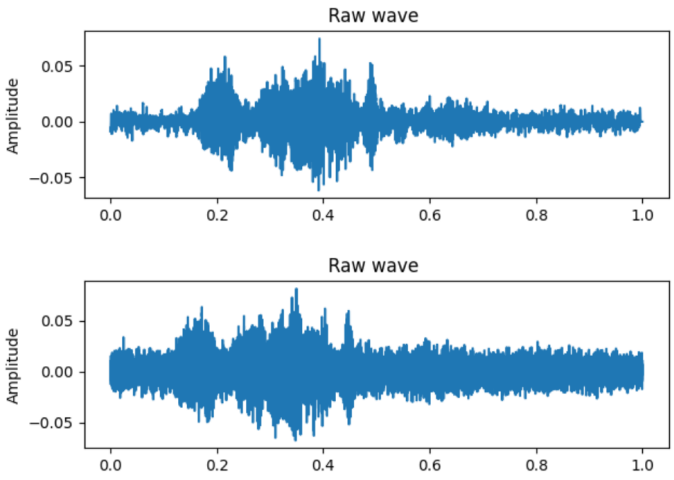

Fig. 5. Image showing the audio wave of the utterance "Tomorrow it will be cold" before and after random shift + noise injection at an SNR of 5

3.3.3 Principal Component Analysis

The final experimentation involved the use of principal component analysis (PCA), which is an unsupervised dimensionality reduction technique. This statistical method works by finding the orthogonal directions (which are called the principal components) that capture the maximum data variance and projecting the data into those directions to create a lower dimensional representation that captures the

majority of information from the data [72]. These principal components are a linear combination of the original variables in the data [72]. The first principal component captures the most amount of variance in the data followed by the next principal component and so on (a visualisation of the convergence can be seen in Fig. 6). The method has proven to be effective with SVMs, given that the features in the data are correlated.

This technique was applied to the data with the aim of removing noise that could have resulted in the model overfitting. The number of principal components that capture 99%, 90% and 80% of the training data's variability were selected to keep information loss to a minimum.

The model trained and tested on data that has not been transformed by PCA is referred to as 'original' in this subsection, and networks trained and tested on data that has been transformed by PCA whilst keeping 99%, 90% and 80% explained variance is referred to as PCA_99, PCA_90 and PCA_80 respectively.

Applying PCA and keeping 99% explained variance reduced the number of features from 6374 to 1740, 90% variance reduced it to 717 features and 80% variance reduced it further to 364 features.

The PCA_90 model achieved the highest cross-validation accuracy of 79.99%, followed by PCA_99, followed by the original model and finally the PCA_80 model which achieved a slightly lower accuracy of 78.46%. The results of the different PCA methods on both 'matching' and 'non-matching' audio-only test sets can be seen in Table 4. Due to its superior performance, the PCA_90 model was selected as the final model used for comparison against humans on audio-only stimuli. This model will be referred to as the SVM model from section 4 onward.

As well as the improvement in accuracy, the significant improvement in computational performance (see Table 5) as a result of applying PCA should also be noted. For example, as well as achieving the highest accuracy on the matching test set, the PCA_90 model generated predictions on this set in 0.746 seconds compared to the original model which took 11.5 seconds (i.e. over 15 times quicker). This can be particularly advantageous for models that are used for real-time emotion detection.

For reproducibility, the PCA_90 model used the values of 100 and 1e-4 for the C and gamma hypermarameters respectively.

TABLE 4
Table showing the accuracies of the model trained on data transformed via PCA whilst keeping a certain amount of data variability (denoted by the number after 'PCA'). CV Acc. means average cross-validation accuracy across the 10 folds.

	CV Acc.	Matching Acc.	Non-matching Acc.
original	79.06	80.32	33.43
PCA_99	79.31	80.48	33.36
PCA_90	**79.99**	**80.97**	34.26
PCA_80	78.46	80.00	**34.56**

However, despite the careful hyperparameter tuning and the various experiments carried out in this section, the SVM model was still achieving 100% training accuracy and the problem of overfitting was not resolved. Thus it was concluded that the overfitting was occurring because the SVM model was able to easily separate the data points in the training set due to the lack of variability of the data.

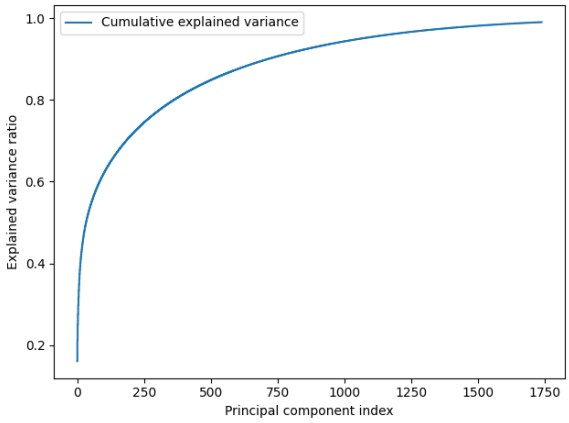

Fig. 6. Curve showing the explained variance of the training set against the number of principal components.

TABLE 5
Time taken (s) to generate predictions on both 'matching' and 'non-matching' audio test sets using the full ComParE_2016 feature set versus the feature set after PCA dimensionality reduction.

	Matching test set	Non-matching test set
original	11.5	80.0
PCA_99	2.61	18.4
PCA_90	0.746	5.65
PCA_80	0.413	2.71

3.4 Visual classification pipeline

3.4.1 Visual Preprocessing

Following the research performed in section 2, a CNN-LSTM architecture proposed by Hans and Rao [47] was proven to show compelling results on the CREMA-D dataset. As such, their approach and network architecture were used as an inspiration for this work.

As suggested by Hans and Rao [47] face frames of size 28x28 pixels were extracted from each .flv file of the visual stimuli using the OpenFace 2.0 toolkit [73] (and a custom script to iterate through each stimulus) and placed in individual directories. The OpenFace toolkit used the MTCNN face detection algorithm to detect the face in each frame, followed by an affine warp transformation to align the face and then performed a crop to remove unwanted detail [73].

Unlike the audio data, SMOTE oversampling was not considered appropriate for balancing the imbalance of visual stimuli. As suggested by Dablain et al. [74], SMOTE-based methods are not only computationally expensive as they require the full image dataset during training and inference, but they also do not perform well on noisy data such as images. Instead, a random oversampling of the minority classes was applied to the training set. The 'RandomOverSampler' method provided by the 'imblearn' library [68] was used to randomly select (with replacement) and oversample the minority class. This resulted in the training data containing 851 instances of each emotion class.

Following this, a custom PyTorch DataLoader [75] was created to load the frames into memory for processing and to extract 75 equally spaced face-aligned frames from each visual stimuli directory. As suggested by Hans and Rao [47], the last frame was duplicated for clips with less than 75 frames. Each stimuli was stored in a 75*3*28*28 tensor along with the class label. Fig. 7 shows a sample of 75 face-aligned frames from a video portraying the emotion of 'anger'.

Sample instances of the train, validation and test DataLoaders [75] were initiated and were iterated through to calculate the mean and standard deviation of each. Once computed, a second set of DataLoaders [75] were created which normalised the RGB channels of each image in the DataLoader [75] with the pre-computed mean and standard deviation. This is a popular step in image processing to help ensure each pixel has a similar data distribution, thus allowing faster convergence when training the model.

Although transformations such as random vertical flips were considered as a pre-processing step for the training data, they were not chosen because they were not appropriate for the context of data used (e.g. frowning lips inverted could resemble a smile, both of which convey different emotions). Random cropping, another popular data augmentation method, was also avoided because the images were already cropped to align and fit the face, so cropping the image further could result in the loss of valuable information. Instead, random horizontal flips with 50% probability were applied to the training batches to reduce the effects of over-fitting and to diversify the oversampled instances.

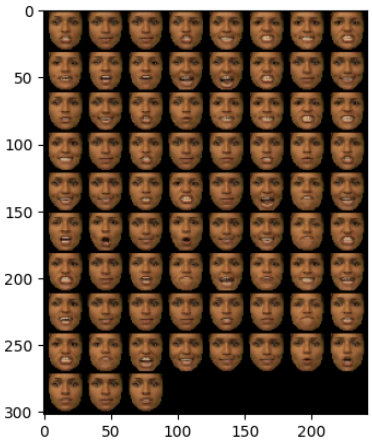

Fig. 7. An example of 75 frames extracted from a visual stimulus portraying the emotion of 'anger'. The frames are plotted on a 2D grid for visualisation.

3.4.2 Model Architecture

Hans and Rao [47] proposed three different CNN-LSTM architectures consisting of different layers. Although their '6-layer' CNN-LSTM model (referred to as CNN-LSTM-2 in our project) performed the best on the CREMA-D dataset, this project also implemented the smaller CNN-LSTM model (referred to as CNN-LSTM-1 in our project) as well. This was because we were working with a much smaller set of data, and simpler model was thought to have the possibility of achieving better performance and show more promising signs of learning. Since we implemented these models from scratch in Pytorch unlike Hans and Rao [47] who used TensorFlow, there were minor implementation differences.

The proposed CNN-LSTM-1 model was constructed to receive an input of tensors of size 75x3x28x28 (75 frames of 28x28 RGB images) which was fed through a convolution block containing: a 3D convolution filter (with a kernel size of 3x3x3), leaky ReLU activation function, 3D batch normalisation and 3D dropout with a probability of 0.07 in that order. The output of the convolution block was passed through an LSTM layer with a dropout rate of 0.14 and a hidden size of 64. Next, the output passed through a fully connected linear layer with 32 units followed by a leaky ReLU activation function and dropout layer with probability 0.21, before finally passing through another fully connected linear layer and a softmax output.

CNN-LSTM-2 has a very similar architecture, with the main difference being the addition of an extra convolution block, which is followed by a 3D max pooling layer (with kernel size 2x2x2) and another convolution block in-between the convolution block and LSTM layer of CNN-LSTM-1.

3.4.3 Improvements to CNN-LSTM architecture

To improve the performance of the CNN-LSTM architectures proposed by Hans and Rao [47], the following changes were made.

The first modification addressed the exploding gradients issue, which is a common problem with models, like the CNN-LSTM network, containing recurrent layers such as RNNs or LSTMs [76].

LSTMs have an advantage over traditional recurrent neural networks (RNNs) by addressing the vanishing gradient problem, which occurs during backpropagation. They do so using a series of input, output and forget gates. This allows the network to selectively retain and forget information as needed for forward propagation, and similarly control the flow of gradients during backpropagation.

[77]. However, exploding gradients is still an issue for LSTM networks. Due to the architectural design, if the values in a previous cell state are large, the gradients resulting from the backpropagation process can grow excessively as well, leading to exploding gradients. This is an issue for training because it can introduce instability in the training process. More specifically, if the gradients during backpropagation are excessively large, this can cause the updates to the weights and biases to be large as well and cause the model to diverge from the optimal solution [77].

To reduce the effects of exploding gradients, the norm of the overall gradient was clipped via gradient clipping. More formally, if the norm of the gradient exceeded the threshold value (set to 1 as is popular in the literature), the gradients were clipped to a value that was calculated by multiplying the unit vector of the gradients with the threshold:

$$\hat{g} \leftarrow threshold \cdot \frac{\hat{g}}{\|\hat{g}\|} \quad if \ \|\hat{g}\| \geq threshold \tag{4}$$

Where $\hat{g}$ is the gradient.

This ensured that the gradients did not exceed a predefined maximum, which consequently led to more stability in training and to the divergence of a more optimal solution.

TABLE 6
A table showing the training results of CNN-LSTM-1 and CNN-LSTM-2 for different learning rates. The epoch in which this maximum validation accuracy was achieved for each model is shown.

Model	Train Acc.	Train Loss	Val. Acc.	Val Loss	lr	Epoch
CNN-LSTM-1	25.68	1.678	29.86	1.661	1e-5	94
CNN-LSTM-1	26.58	1.636	30.68	1.612	1e-4	65
CNN-LSTM-1	16.20	1.791	23.51	1.782	1e-3	73
CNN-LSTM-2	89.84	0.3169	80.95	0.7240	1e-5	125
CNN-LSTM-2	**96.28**	**0.1465**	**82.97**	**0.4785**	1e-4	114
CNN-LSTM-2	25.38	1.672	31.62	1.595	1e-3	93

Secondly, since the Pytorch implementation of cross-entropy loss (used as the criterion for training) applies a softmax on the output of the network followed by a negative log-likelihood loss [75], the final softmax layer was removed from CNN-LSTM-1 and CNN-LSTM-2 due to redundancy.

These changes together resulted in over 10% increase in the accuracy of the best-performing model identified in the next section. Fig. 8 outlines the modified architectures.

3.4.4 Model Training

Both models were trained on the visual data with a batch size of 4, and as suggested by Hans and Rao [47], an Adam optimiser with beta values of 0.99 and 0.999 and loss function using cross-entropy loss. Each model was trained with learning rates of 1e-3, 1e-4 and 1e-5. Although Hans and Rao [47] only trained the models for a maximum of 75 epochs, this project trained both for 100 epochs since the size of the dataset used was much smaller.

After training for the first 100 epochs, the CNN-LSTM-2 network trained with learning rates of 1e-4 and 1e-5 continued to show signs of learning as their accuracies were steadily increasing while the loss continued to diminish. Therefore both networks were trained for a total of 150 epochs (see the training progression of the CNN-LSTM-2 model in Fig. 9). The results are shown in Table 6. Due to its superior performance, the CNN-LSTM-2 network with a learning rate of 1e-4 was selected as the visual model for comparison with human data. This model will be referred to as the CNN-LSTM model from section 4 onward.

3.4.5 Attempt to reduce Overfitting

Attempts were made to reduce overfitting by introducing a weight decay term to the Adam optimiser. Although a range of weight decay terms was used to minimise the difference between the training and validation accuracy, the reduction in validation accuracy was too significant and so this approach was discontinued.

3.5 Evaluation metrics

As well as comparing the accuracy of each model against humans, the precision (eq. 5), recall (eq. 6) and f1-score (eq. 7) were calculated on a per-class and macro-averaged (better suited than weighted-average for imbalanced classes as classes are treat equally) level for both test sets. The precision score depicts how many of the predictions for a given class were accurate, whereas the recall score provides an understanding of how many of the class instances were identified. The f1-score calculates a harmonic mean between precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{7}$$

These metrics were chosen not only because they're popular for classification, but they allow us to answer the research question from multiple angles which may be more desired in certain domains. For example, recall is important in domains where the number of false negatives (FN) should be minimised, precision in domains where the chances of predicting false positives (FP) are minimised and the F1-score is preferred over accuracy when the dataset has a class imbalance (which is the case for our test sets).

These metrics were computed for both ML models and humans for comparison. The scores calculated for humans contained the responses from all human raters who gave ratings for items in the respective test set e.g., the human precision score for the matching test set took the true positives and false positives from all the individual rater responses in that set into consideration.

4 Results

4.1 Overall Results

Both the CNN-LSTM and SVM models outperformed humans for all measured macro metrics on both the matching and non-matching test sets as seen in Table 7 and Table 8.

For both audio and visual data, although the models suffered a significant drop in performance when generating predictions for the non-matching test set (as seen for humans), the difference in the macro-averaged evaluation metrics on both audio and visual data between humans and the ML models were much greater than the difference seen for the matching test set. In fact, the ML models used for both audio and visual data achieved approximately double the accuracy as that seen for humans on the non-matching set. This could suggest that ML models can capture relevant information from stimuli even under more ambiguous conditions to deliver more accurate predictions than humans.

Looking at the results on a granular, individual emotion class level, the results (see Fig. 10, Fig. 11, Fig. 12, and Fig. 13) showed that although ML models achieved higher performance scores for the majority of classes, humans had a slight advantage in some emotion categories. For example, for the visual matching data, the CNN-LSTM model achieved a 66% accuracy (i.e. recall score) in identifying the 'anger' emotion, whereas humans performed much greater with a 10% higher accuracy (although the CNN-LSTM model achieved greater precision by 0.02%). This could be a result of relevant features in the matching test set for the 'anger' emotion did not appear or weren't as relevant in the training data the ML model was given. On the other hand, it may alternatively suggest that humans may have an advantage over ML models at classifying a subset of stimuli from certain emotion classes like 'anger'. Nevertheless, the ML model had a significantly higher accuracy of 35% compared

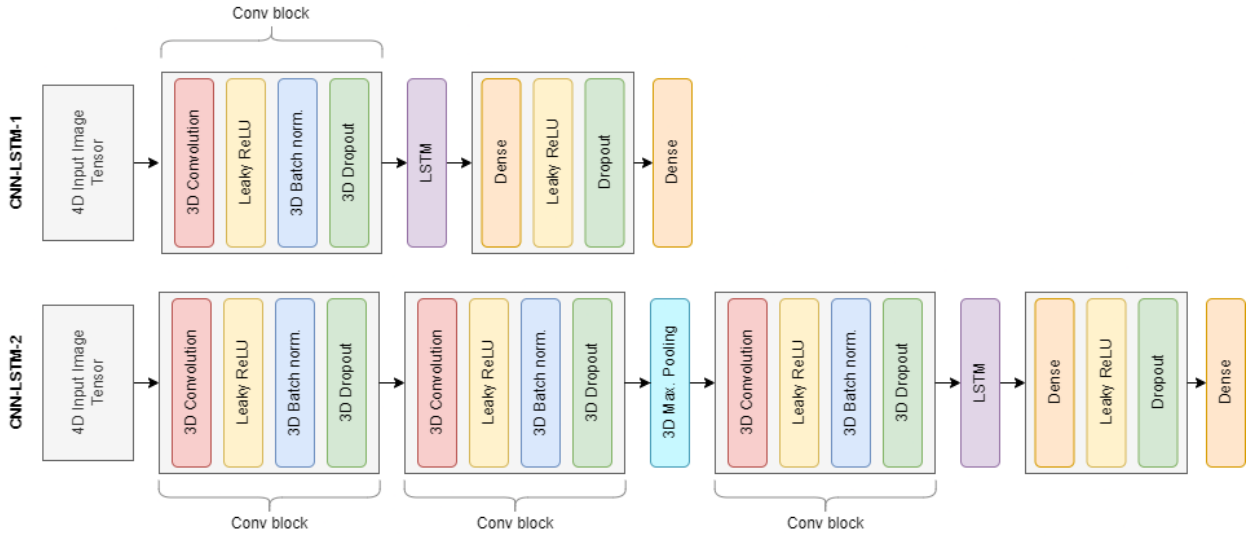

Fig. 8. Figure showing the architecture of CNN-LSTM-1 and CNN-LSTM-2.

TABLE 7
Table showing the macro-averaged results for the CNN-LSTM model and humans on the matching and non-matching visual test set.

Classifier	Matching test set				Non-matching test set			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
CNN-LSTM	**80.81**	**78.93**	**77.97**	**78.23**	**38.96**	**39.96**	**43.29**	**35.89**
Humans	75.91	73.53	73.53	73.40	19.37	23.11	22.07	19.47

TABLE 8
Table showing the macro-averaged results for the SVM model and humans on the matching and non-matching audio test set.

Classifier	Matching test set				Non-matching test set			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
SVM	**80.97**	**77.45**	**73.85**	**75.34**	**34.26**	**45.31**	**44.44**	**33.40**
Humans	68.91	65.72	64.90	64.82	17.90	31.56	20.51	19.90

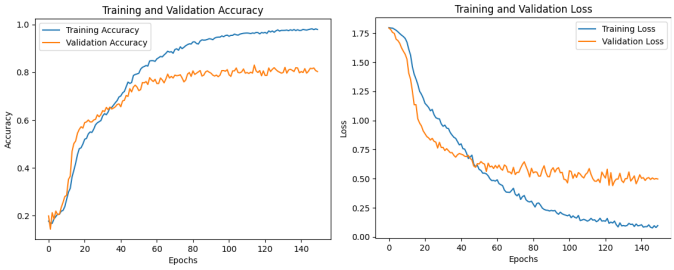

Fig. 9. Graphs showing the accuracy and loss for CNN-LSTM-2 on both the training and validation set for 150 epochs.

to the 21% averaged accuracy of human raters for the 'anger' emotion on the non-matching test set.

A similar pattern was seen for the precision score for the emotion of 'fear', where on the matching visual test set, the CNN-LSTM model achieved a precision score of 64%. In contrast, the average precision of human raters was 73%. This implies that humans made fewer classifications when the intended emotion for the stimuli was 'fear'. However, as seen before, the ML model outperformed humans on the non-matching test set for the emotion of fear on all measured metrics including precision, where the ML model achieved a precision score of 42% whereas humans achieved a lower score of 34%. This again supports the claim that ML models may perform better on more ambiguous data.

For the visual non-matching test set, the only emotion in which humans achieved a higher score than the ML model was 'happy', for which the ML model achieved a precision score of 11% and the human average was 14%. This led to humans also achieving a higher f1-score than the ML model on the emotion of 'happy'. The superior ability of humans to achieve a lower number of false positives for the emotion categories 'happy' and 'fear' (as indicated above) was also observed in the matching test set.

For the audio data, the SVM model achieved a higher accuracy than humans for all emotion classes in the matching test set. The ML model also achieved a higher accuracy score for all classes in the non-matching test set as well, except for when the intended emotion was 'sad', where the ML model achieved a recall score of 17% whereas humans achieved a marginally higher accuracy of 18%.

Again, similar to the visual test set results, although the ML model has greater precision than humans for the majority of emotion classes, humans were seen to achieve a higher precision score of 79% compared to the 70% achieved by the ML model for the 'happy' emotion class on the matching audio test set. This consequently led to humans also achieving an f1-score that was 0.02% higher than the

ML model for the 'happy' emotion class. Humans also achieved a fractionally higher precision score (less than 1%) for the 'anger' emotion class on the matching test set.

The greater precision scores for humans suggest that humans may be less prone to incorrectly classifying emotions that don't belong to a class as belonging to it. In summary, Our data suggest that these emotions include 'happy' and 'fear' for the visual data, and 'happy' and 'angry' for the audio data. Nevertheless, it should also be noted that the difference between the two classifier precision scores are very minute and the lower performance of the ML models may be overcome with more training on a varied set of data.

4.2 Misclassifications of the 'neutral' emotion class

To answer the sub-research question exploring the 'neutral' emotion, charts (see Fig. 14 and Fig. 15) were made to plot the percentage of ratings for each class that were misclassified as 'neutral' (out of the total number of ratings for that class).

For all audio and visual sets, the emotion most commonly misclassified by humans as 'neutral' was 'sad'. As seen in Fig. 14, 22% of audio-only stimuli from the matching test set and 59% from the non-matching test set were misclassified as neutral. Interestingly, this phenomenon was also observed for the SVM model which also had the highest percentage of 'neutral' misclassifications when the intended emotion was 'sad' (with only 3% fewer misclassifications than humans on both test sets). As a result of the vast number of misclassifications, the precision score for humans and the SVM model for the 'neutral' emotion class on the non-matching test set is substantially low (as seen in Fig. 11, sitting at 1% and 3% respectively. This may suggest that audio data in the test sets which portray the emotion of 'sad' and 'neutral' both have very similar audio features, making them difficult to distinguish. However, it should also be noted that the audio training set had a very limited number of samples for the 'sad' emotion class (see Fig. 3). Thus, although SMOTE oversampling was performed to create synthetic samples of the minority class, the poor performance of the ML model may have also been a result of the lack of diversity in the 'sad' audio training data.

For the visual data on the other hand, this pattern is not observed and as shown in Fig. 15. Although the CNN-LSTM misclassified 32% of 'sad' visual stimuli as 'neutral' on the non-matching test set (which is still a substantial amount), this is significantly less than humans who misclassified 54% of the stimuli.

It should also be noted that for both audio and visual data, the ML models had a significantly lower 'neutral' misclassification rate for stimuli belonging to the 'happy' class. For example, on the non-matching audio test set, the SVM model misclassified 15% of 'happy' audio clips as 'neutral' compared to humans who exhibited a 53% misclassification rate. Similarly, for the non-matching visual test set, the CNN-LSTM model misclassified 16% of the 'happy' visual data as 'neutral' compared to humans who misclassified 47%.

Additionally, for emotions where humans had the lowest percentage of 'neutral' misclassifications ('anger' for audio and 'disgust' for visual), it was found that ML models were able to achieve almost perfect results with less than 5% 'neutral' misclassifications on both matching and non-matching test sets. In fact, the SVM classifier made no misclassifications of the 'anger' stimuli as belonging to the 'neutral' class. This shows that the ML models were able to learn information that provides a strong separation between these said emotions and 'neutral', particularly those emotions for which humans made fewer misclassifications (possibly because these have separable features that ML models can identify more consistently than humans).

The only emotion for which humans exhibited a smaller percentage of 'neutral' misclassifications was 'anger' on the visual matching test set. On this set, it was found that the ML model performed much worse, yielding a 9% higher 'neutral' misclassification rate for 'anger' stimuli (this was the most commonly misclassified emotion by the model on the matching set). Although the CNN-LSTM model did achieve a lower misclassification rate on the non-matching test set, the percentage difference in misclassifications by humans and the ML model (2%) was negligible.

In summary, both SVM and CNN-LSTM models outperformed humans in classifying emotions, with a lower percentage of 'neutral' misclassifications (except the CNN-LSTM model for the 'anger' class on the matching test set). Further, when humans had a low percentage of 'neutral' misclassifications, the ML models performed even better. Finally, although the SVM model yielded a similar percentage of 'neutral' misclassifications for 'sad' audio clips, this may have been a result of the lack of variety in training data.

5 Evaluation

5.1 Suitability of approach

By researching SOTA ML models in the domain of SER and FER, we were able to implement a well-performing model for each modality, answering the research question with the conclusion that it is possible for ML models to recognise most emotions better than humans (given the assumptions of this study).

Further, by splitting the data set into a matching and non-matching test set, we took an extra step to back up our claim and show that ML models can have remarkable performance, particularly when humans have difficulty distinguishing different emotions.

Finally, by evaluating the performance of humans and the ML models on a per-emotion basis, not only were we able to analyse which emotions the classifiers performed well in, but we also were able to explore our sub-research question of whether models tend to misclassify a large percentage of items as 'neutral' as seen for humans and report our findings.

5.2 Strengths

The chosen methodology and approach have several strengths. For example, careful consideration was given to choosing an appropriate model based on the context and lack of training data available. This was a particular issue for the audio data as a large portion of the data belonged in the non-matching test set (Appendix Fig. 16), so a choice was made to use an SVM model instead of a CNN-based

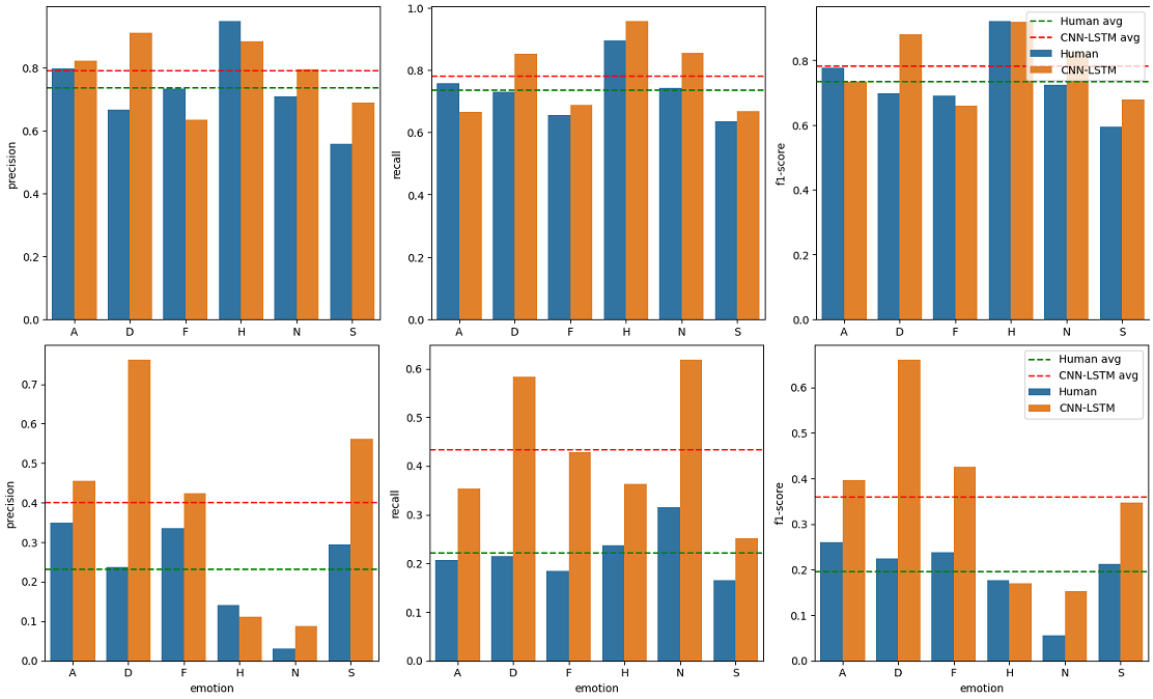

Fig. 10. Charts showing the results of the CNN-LSTM model against humans for precision, recall and f1-score. The first row of charts shows the results on the matching visual test set and the second row on the non-matching visual test set.

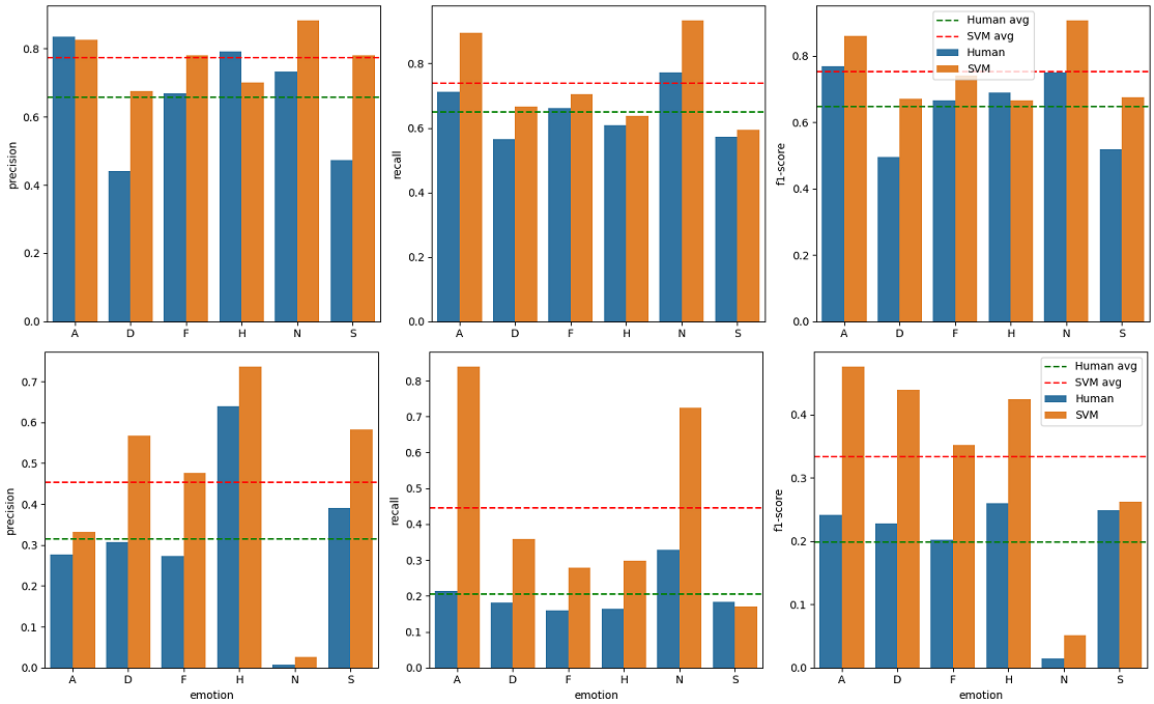

Fig. 11. Charts showing the results of the SVM model against humans for precision, recall and f1-score. The first row of charts shows the results of the matching audio test set and the second row on the non-matching audio test set.

architecture (which requires large amounts of data for good results). This proved to work out very well and high levels of performance were achieved by the model.

To maintain a rigorous scientific approach, we enforced good practices by keeping training, validation and testing sets independent to prevent data leakage during training. For example, for the visual classification pipeline, this involved calculating the mean and standard deviation for each of the train, validation and test data individually before normalising the data.

A further strength of this project was the vast amount of optimisations and experiments that were carried out to improve the ML models, and extended beyond just the hyperparameter tuning. For the SVM model, this involved

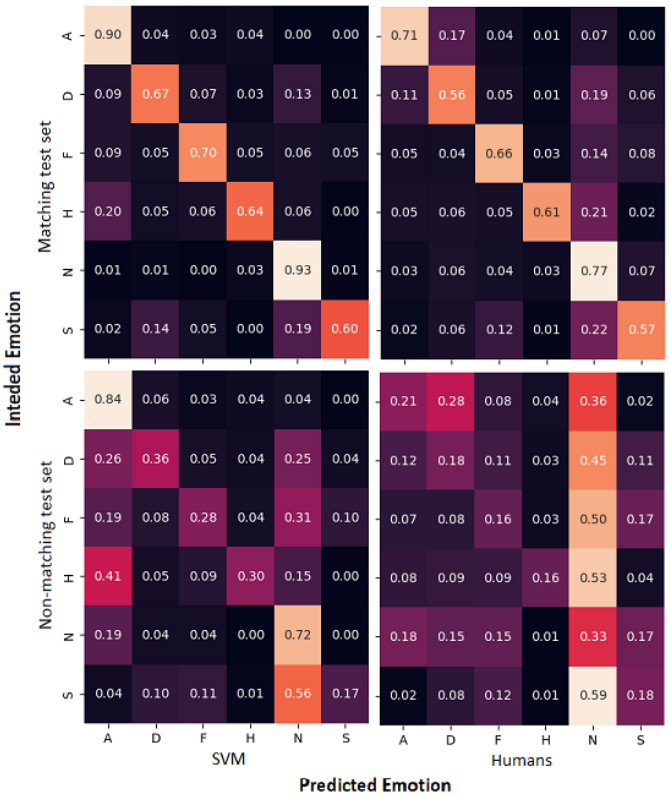

Fig. 12. Confusion matrix comparing SVM model and human performance on matching and non-matching audio tests. Black represents 0% and white represents 100%. Purple, red, and orange are intermediate ascending values. The colour map was adjusted for each matrix to emphasise important areas.

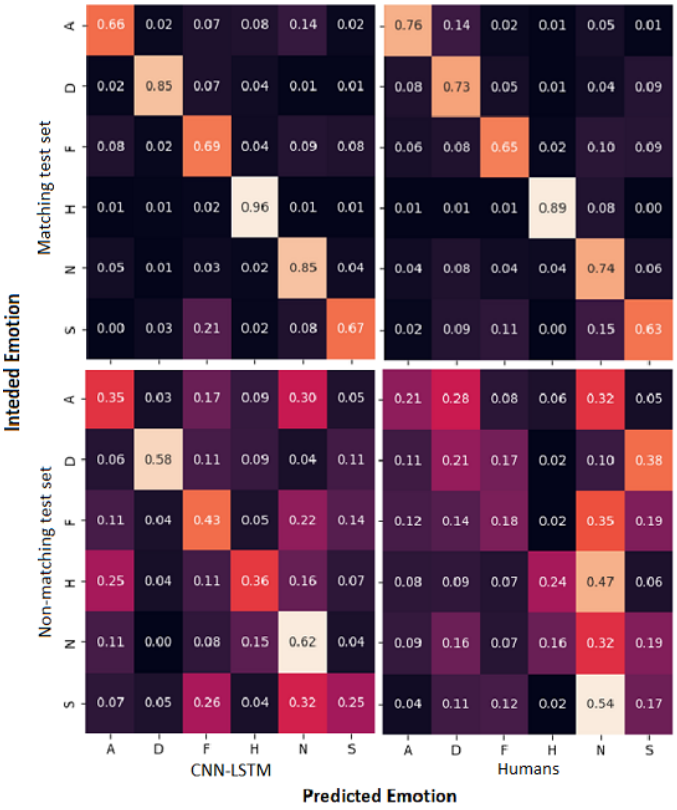

Fig. 13. Confusion matrix comparing CNN-LSTM model and human performance on matching and non-matching audio tests. Black represents 0% and white represents 100%. Purple, red, and orange are intermediate ascending values. The colour map was adjusted for each matrix to emphasise important areas.

experimenting with different audio feature sets, experimenting with data augmentation, and applying different levels of dimensionality reduction. For the CNN-LSTM model, this included experimenting with two different architectures, applying gradient clipping, modifying the output layer, and applying weight decay to reduce overfitting. These improvements not only enhanced the model performance but also highlighted the secondary benefits of such methods. For example, applying PCA to the SVM model not only improved its accuracy but also substantially decreased the computational cost by over 14-fold. We hope that the successful experiments, as well as those that were unsuccessful, act as a guide for future researchers in this area.

Finally, to evaluate the research question from multiple angles, the performance of humans and the ML models were analysed using a range of popular metrics including precision and recall. Since looking at the accuracy alone can be deceiving, using these metrics gave a better insight into which areas resulted in poor performance. This can be very useful when considering the use of ML algorithms for emotion detection in specific domains.

5.3 Limitations of methodology

Due to the nature of this project, a key limitation was the lack of ability to verify the results.

Firstly, as reported in section 2, no other dataset was found to be a suitable candidate to evaluate the performance of the trained ML models against humans. This was primarily because no other dataset provided granular

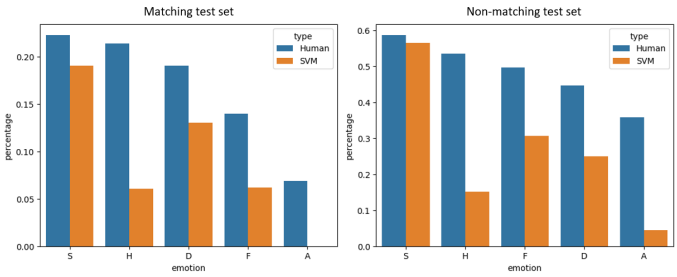

Fig. 14. Charts showing the percentage of misclassification made by humans and the SVM model on the matching and non-matching audio test sets.

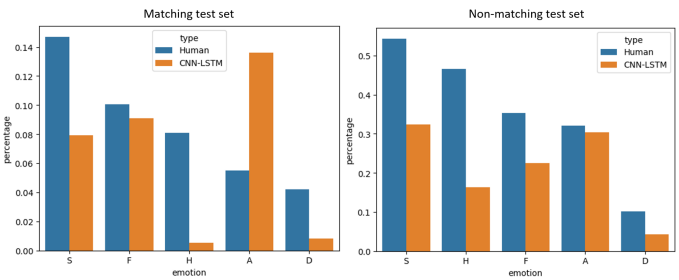

Fig. 15. Charts showing the percentage of misclassification made by humans and the CNN-LSTM model on the matching and non-matching visual test sets.

human rater results for each stimulus. This meant that there was no opportunity to verify if the observed results occur on other datasets collected from different environments. Further, the ML models were trained on data from actors of which some also appear in the test sets (the limited data availability prevented us from maintaining unique actors for both samples), which may also have given the ML models an advantage on the CREMA-D dataset.

Secondly, the use of statistical significance tests to compare human and ML performance was not suitable due to the limited data. For example, we considered conducting a Student's t-test with the null hypothesis being that the mean of human and ML model accuracy are equal. However, this approach was not feasible because n samples of human data were required to conduct the test. By the definition of our non-matching test set, it was not possible to have more than one test set as this contained all the stimuli that fall into the non-matching and ambiguous category as originally defined by Cao et al. [18]. Alternatively, splitting the test set would result in too few occurrences of some emotions.

Although a one-sample t-test was considered to overcome the first issue (i.e. using a single human score value), t-tests require n independent train and test splits that do not overlap [78]. This was problematic because although it would be possible to generate n training sets through random re-samples of the full training set, this would lead to a high type 1 error (i.e. the null hypothesis is incorrectly rejected) [78]. Due to the lack of data in some classes, splitting the training data into n independent sets was not feasible either. Since the assumption of independence would be violated, this hypothesis test was not used. For similar reasons, other popular tests such as the Wilcoxon signed rank test were also inappropriate.

An alternative approach to comparing the performance of ML models to humans would be to collect our own data and follow the approach of Goh et al. [79] who compared human classifiers (i.e. students) to ML models by training and testing both human and ML classifiers on the same data. They conducted the training phase for human classifiers by giving them the training data abstract each morning and telling them the ground truth. By training the ML models and human classifiers on the same independent training and test data, they were also able to conduct hypothesis testing using a two-sided t-test. However, given the time constraints and complexity of conducting a large-scale data collection phase, this was too difficult to do for our project.

Finally, overfitting was an issue that occurred for both ML models, but more particularly for the audio ML model. Although several techniques were employed to reduce this phenomenon, the issue was not resolved for both audio and visual ML models. The likely underlying cause was attributed to the lack of diversity in the data, posing a significant limitation for these models in terms of their ability to generalise and perform well on unseen data.

5.4 Discussion of algorithms used

To summarise, the techniques chosen for this project performed well for the given task. It was evident that through the rigorous optimisation and experimentation carried out in this paper, the SOTA ML models produced were able to outperform human raters on all selected macro-averaged metrics, and also on most individual emotion classes.

6 Conclusion

To conclude, this project demonstrated that it is possible for ML models to recognise emotions better than humans as illustrated by the superior performance of both models over humans on accuracy and all the macro-averaged metrics (precision, recall and f1-score). We also demonstrated the advantage of ML models when given data that humans have difficulty discerning, not only by showing the increased difference in performance between ML models and humans on the non-matching test set where the ML models achieved almost double the accuracy, but also by showing that ML models make fewer misclassifications (particularly focusing on 'neutral'). It was observed that for emotions for which humans achieved the lowest percentage of 'neutral' misclassifications, the ML models performed even better as they were able to consistently separate the classes from 'neutral'. For example, the SVM model made no misclassifications of the 'anger' emotion stimuli from the matching test set as belonging to 'neutral' as seen in Fig. 14.

However, this does come with a caveat as humans were seen to have better performance for some emotion categories when looking at a granular level. For example, humans had a 10% higher accuracy than ML models for capturing the emotion of 'anger' from visual data on the matching test set. We also saw that humans achieved slightly higher precision scores on one or both of the tests for the emotions of 'happy' and 'fear' for the visual data, and 'happy' and 'angry' for the audio data. We also observed that some emotions, like 'anger' for the visual stimuli, resulted in an unexpectedly large percentage of 'neutral' misclassifications from the ML model on the matching test set. It is possible that these could be a result of the lack of variety in training data.

We hope that these findings, along with the granular analysis, provide future researchers with a greater understanding of the performance of ML models against humans in the domain of emotion classification. We hope that the range of metrics used for the evaluation can help make judgements on the suitability of ML models to cater to different business needs e.g. the use of ML models in domains where greater precision is required for specific emotions.

In addition to the contributions to the research question, the experimentation performed in this study for audio and visual data showed remarkable improvements in the accuracy and speed of the models. Most notably, applying PCA to reduce the dimensionality of the data that was processed by the SVM resulted in more accurate predictions being generated 14 times quicker.

As a major limitation of this project was the lack of datasets available for comparison, future work should focus on making fine-grained data of individual raters available in datasets, as seen in the CREMA-D dataset. This enables the verifiability of the results of similar work. Additionally, future work comparing ML models to humans should undertake a similar approach to Goh et al. [79], who trained and tested human classifiers on the same subsets as ML models. This method allows statistical significance tests to be carried out for a more rigorous evaluation.

APPENDIX

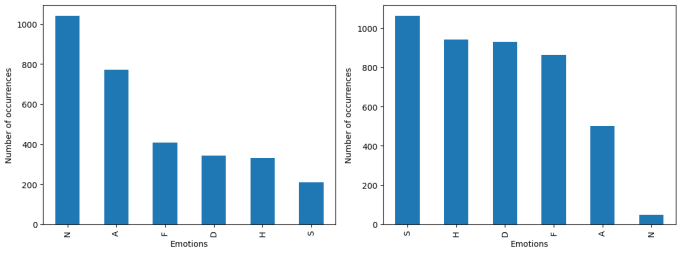

Fig. 16. Charts showing the number of emotion class instances in the audio-only matching and non-matching sets before train/test split. The matching set is on the left, and non-matching is on the right. N=Neutral, A=Anger, F=Fear, D=Disgust, H=Happy, S=Sad.

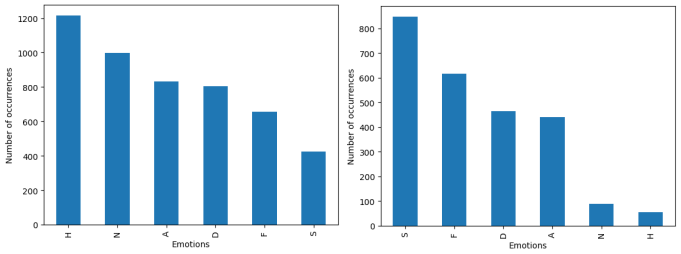

Fig. 17. Charts showing the number of emotion class instances in the video-only matching and non-matching sets before train/test split. The matching set is on the left, and non-matching is on the right. N=Neutral, A=Anger, F=Fear, D=Disgust, H=Happy, S=Sad.

REFERENCES

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300738

[2] R. Picard, "Affective computing. 321," 1995.

[3] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 1034–1041.

[4] K. O. Akputu, K. P. Seng, and Y. L. Lee, "Facial emotion recognition for intelligent tutoring environment."

[5] P. Ekman, "Universals and cultural differences in facial expressions of emotion." in *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[6] G. Brodny, A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions," in *2016 9th International Conference on Human System Interactions (HSI)*. IEEE, 2016, pp. 397–404.

[7] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *Journal of nonverbal behavior*, vol. 43, pp. 133–160, 2019.

[8] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, "The relation between valence and arousal in subjective experience." *Psychological bulletin*, vol. 139, no. 4, p. 917, 2013.

[9] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[10] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *The Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 88–98, 1995.

[11] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[12] A. Protopapas and P. Lieberman, "Fundamental frequency of phonation and perceived emotional stress," *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2267–2277, 1997.

[13] M. G. Beaupré and U. Hess, "Cross-cultural emotion recognition among canadian ethnic groups," *Journal of cross-cultural psychology*, vol. 36, no. 3, pp. 355–370, 2005.

[14] D. Keltner and D. T. Cordaro, "Understanding multimodal emotional expressions," *The science of facial expression*, 1798.

[15] J.-M. Fernández-Dols and M.-A. Ruiz-Belda, "Are smiles a sign of happiness? gold medal winners at the olympic games." *Journal of personality and social psychology*, vol. 69, no. 6, p. 1113, 1995.

[16] I. Eibl-Eibesfeldt, "11. similarities and differences between cultures in expressive movements," *Non-verbal communication*, p. 297, 1972.

[17] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures." *Emotion*, vol. 18, no. 1, p. 75, 2018.

[18] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[19] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[20] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[21] D. C. Ong, Z. Wu, Z.-X. Tan, M. Reddan, I. Kahhale, A. Mattek, and J. Zaki, "Modeling emotion in complex stories: the stanford emotional narratives dataset," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 579–594, 2019.

[22] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[23] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, mar 2021.

[24] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*. IEEE, 2008, pp. 865–868.

[25] E. Douglas-Cowie, C. Cox, J.-C. Martin, L. Devillers, R. Cowie, I. Sneddon, M. McRorie, C. Pelachaud, C. Peters, O. Lowry, A. Batliner, and F. Hoenig, *The HUMAINE database*, 10 2011, pp. 243–284.

[26] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[27] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "Emoreact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th acm international conference on multimodal interaction*, 2016, pp. 137–144.

[28] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[29] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–7.

[30] 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, p. e0196391, 2018.

[31] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[32] S. Haq, P. J. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*, 2008.

[33] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: a chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 913–924, Sep. 2016. [Online]. Available: https://doi.org/10.1007/s12652-016-0406-z

[34] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd international conference on data engineering workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[35] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.

[36] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885617300379

[37] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 973–982.

[38] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, Jan. 2018. [Online]. Available: https://doi.org/10.3390/s18020401

[39] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[40] E. Ryumina, D. Dresvyanskiy, and A. Karpov, "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study," *Neurocomputing*, vol. 514, pp. 435–450, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222012656

[41] J. Chen, Z. Chen, Z. Chi, H. Fu *et al.*, "Facial expression recognition based on facial components detection and hog features," in *International workshops on electrical and computer engineering subfields*, 2014, pp. 884–888.

[42] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on fer2013," *arXiv preprint arXiv:2105.03588*, 2021.

[43] N.-C. Ristea, L. C. Duțu, and A. Radoi, "Emotion recognition system from speech and visual information based on convolutional neural networks," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, pp. 1–6.

[44] G. I. Winata, O. P. Kampman, and P. Fung, "Attention-based lstm for psychological stress detection from spoken language using distant supervision," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6204–6208.

[45] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "Ramas: Russian multimodal corpus of dyadic interaction for studying emotion recognition," 03 2018.

[46] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *CoRR*, vol. abs/1811.07770, 2018. [Online]. Available: http://arxiv.org/abs/1811.07770

[47] A. S. A. Hans and S. Rao, "A cnn-lstm based deep neural networks for facial emotion detection in videos," *International Journal Of Advances In Signal And Image Sciences*, vol. 7, no. 1, pp. 11–20, 2021.

[48] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[49] F. Eyben, M. Wöllmer, and B. Schuller, "Openear—introducing the munich open-source emotion and affect recognition toolkit," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 2009, pp. 1–6.

[50] ——, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[51] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[52] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi, and N. Ismail, "Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks," in *2020 6th International Conference on Wireless and Telematics (ICWT)*. IEEE, Sep. 2020. [Online]. Available: https://doi.org/10.1109/icwt50448.2020.9243622

[53] A. Ivanov and G. Riccardi, "Kolmogorov-smirnov test for feature selection in emotion recognition from speech," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 5125–5128.

[54] M. Gokilavani, H. Katakam, S. A. Basheer, and P. Srinivas, "Ravdness, crema-d, tess based algorithm for emotion recognition using speech," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022, pp. 1625–1631.

[55] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017, advances in Cognitive Engineering Using Neural Networks. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089360801730059X

[56] T. Özseven, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320–326, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X18309915

[57] T. Özseven, "The acoustic cue of fear: investigation of acoustic parameters of speech containing fear," *Archives of Acoustics*, vol. 43, no. 2, pp. 245–251, 2018.

[58] R. Singh, H. Puri, N. Aggarwal, and V. Gupta, "An efficient language-independent acoustic emotion classification system," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3111–3121, Dec. 2019. [Online]. Available: https://doi.org/10.1007/s13369-019-04293-9

[59] D. Valles and R. Matin, "An audio processing approach using ensemble learning for speech-emotion recognition for children with ASD," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, May 2021. [Online]. Available: https://doi.org/10.1109/aiiot52608.2021.9454174

[60] R. Matin, "Developing a speech emotion recognition solution using ensemble learning for children with autism spectrum disorder to help identify human emotions," 2020.

[61] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[62] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "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, 2015.

[63] B. Schuller, S. Steidl, A. Batliner, J. Krajewski, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and S. Schnieder, "The interspeech 2014 computational paralinguistics challenge: Cognitive physical load," 09 2014.

[64] M. Sheikhan, M. Bejani, and D. Gharavian, "Modular neural-svm scheme for speech emotion recognition using anova feature selection method," *Neural Computing and Applications*, vol. 23, pp. 215–227, 2013.

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

[66] M. Kuhn and K. Johnson. Springer, 2016, p. 30–31.

[67] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

[68] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[69] G. James, D. Witten, T. Hastie, and R. Tibshirani, p. 184.

[70] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, 2022.

[71] L. N. Wijayasingha, "Adding noise to audio clips," Jan 2021. [Online]. Available: https://medium.com/analytics-vidhya/adding-noise-to-audio-clips-5d8cee24ccb8
[72] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.
[73] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.
[74] D. Dablain, B. Krawczyk, and N. V. Chawla, "Deepsmote: Fusing deep learning and smote for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[76] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013. [Online]. Available: http://arxiv.org/abs/1308.0850
[77] R. C. Staudemeyer and E. R. Morris, "Understanding lstm–a tutorial into long short-term memory recurrent neural networks," *arXiv preprint arXiv:1909.09586*, 2019.
[78] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
[79] Y. C. Goh, X. Q. Cai, W. Theseira, G. Ko, and K. A. Khor, "Evaluating human versus machine learning performance in classifying research abstracts," *Scientometrics*, vol. 125, no. 2, pp. 1197–1212, Jul. 2020. [Online]. Available: https://doi.org/10.1007/s11192-020-03614-2