An Audio Processing Approach using Ensemble Learning for Speech-Emotion Recognition for Children with ASD

Damian Valles
Ingram School of Engineering
Texas State University
San Marcos, USA
dvalles@txstate.edu

Rezwan Matin
Ingram School of Engineering
Texas State University
San Marcos, USA
r_m727@txstate.edu

***Abstract*— Children with Autism Spectrum Disorder (ASD) find it difficult to detect human emotions in social interactions. A speech emotion recognition system was developed in this work, which aims to help these children to better identify the emotions of their communication partner. The system was developed using machine learning and deep learning techniques. Through the use of ensemble learning, multiple machine learning algorithms were joined to provide a final prediction on the recorded input utterances. The ensemble of models includes a Support Vector Machine (SVM), a Multi-Layer Perceptron (MLP), and a Recurrent Neural Network (RNN). All three models were trained on the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS), the Toronto Emotional Speech Set (TESS), and the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). A fourth dataset was used, which was created by adding background noise to the clean speech files from the datasets previously mentioned. The paper describes the audio processing of the samples, the techniques used to include the background noise, and the feature extraction coefficients considered for the development and training of the models. This study presents the performance evaluation of the individual models to each of the datasets, inclusion of the background noises, and the combination of using all of the samples in all three datasets. The evaluation was made to select optimal hyperparameters configuration of the models to evaluate the performance of the ensemble learning approach through majority voting. The overall performance of the ensemble learning reached a peak accuracy of 66.5%, reaching a higher performance emotion classification accuracy than the MLP model which reached 65.7%.**

Keywords—Ensemble, SVM, MLP, RNN, Autism, emotions

I. INTRODUCTION

Human beings are social creatures. For centuries, interdependence has played a major role in the development of societies, and communication is the foundation of any society. It is through communication that useful information is exchanged, and thoughts and feelings are shared with others. Emotional *valence* is a qualitative measure that helps categorize the different types of human emotions. Valence exists over a spectrum, where emotions such as excitement and joy are classified as emotions with positive valence, while emotions like fear and sadness are considered emotions with negative valence. By detecting the emotional valence in a conversation, the listener can take appropriate actions [1]. Like valence, arousal is another measure that indicates the intensity of an emotion. In a dimensional model for emotions, as the one shown in Fig. 1, valence and arousal are the two dimensions over which emotions are categorized.

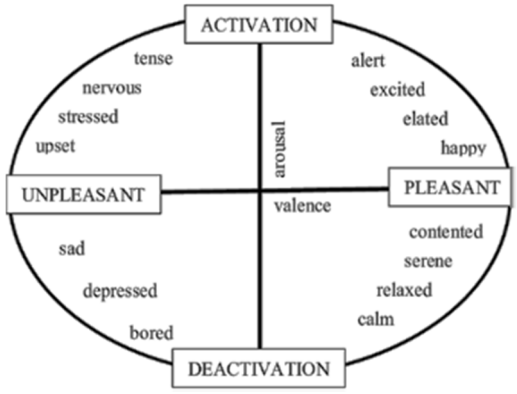

Fig. 1. An example of a dimentional model – The Circumplex Model [2].

Psychologists who follow the dimensional model of emotions state that every emotion can be placed somewhere in these two dimensions. They believe that a standard neurophysiological system generates all human emotions. Other psychologists disagree with this categorization and think that every emotion is created from a different neural system. Since this work follows the discrete categorization of emotions, measuring valence and arousal was not required.

Communication between humans can be either verbal or non-verbal. Non-verbal communication can be through facial expressions or body language. In the year 1971, Albert Mehrabian explained the *7-38-55 rule* of personal communication [3]. The rule states that, in any social interaction, 55% of the information being conveyed comes from the speakers' body language, 38% comes from their vocal tone, and only 7% comes from the spoken words. Also in that year, the researchers in [4] listed the six universally recognized emotions observed across all cultures around the world. This list of emotions includes happiness, sadness, surprise, anger, fear, and

2021 IEEE World AI IoT Congress (AIIoT) | 978-1-6654-3568-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/AIIOT52608.2021.9454174

disgust. For this research work, the six affective states listed in [4] were considered, with the addition of the neutral emotion.

The objective of this work is to summarize the audio datasets that were utilized for the development, training, and analysis of the Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Recurrent Neural Network (RNN) models. Additionally, each model used in the approach is discussed along with the results obtained in classifying each of the discrete emotions. Furthermore, the audio processing of the audio sample files is detailed as it pertains to each of the models' training and analysis. It also explains the audio feature extraction performed on the samples by using Mel-Frequency Cepstral coefficients (MFCCs), Spectral constant, Polynomial coefficients, and Root-Mean-Square (RMS) Energy. Finally, it discusses the approach taken and results obtained when executing the discrete emotions' classification on an ensemble learning approach using a majority voting technique.

II. Relevant Work

The growing interest in the use of machine learning with speech processing can be linked to the massive improvements in computational power over the past few years. A few machine learning and deep learning classifiers in particular have proven to yield excellent results in speech emotion recognition results. The three most common algorithms are the SVM, the MLP, and the RNN. The researchers in [5] used binary SVM models on three different layers for performing multi-class emotion recognition. Each model was trained on a single emotion and classified that emotion against the other emotions in a one-versus-all (OVA) fashion. The dataset used is called the Interactive Emotional Dyadic Motion Capture (IEMOCAP). The Kaldi toolkit was used to extract frame-based features like energy, pitch, MFCCs, perceptual linear predictive (PLP), a filter bank, and the first and second derivatives of all features.

In [6], a multi-lingual speech emotion classifier was built by training an MLP on the Emotional Prosody Speech and Transcripts (EPST), a speech emotion corpus in the English language, and the KSUEmotions, an Arabic speech emotion corpus.. Audio features such as pitch, intensity, formants, jitter, shimmer, and speech rate were used. These features were extracted using PRAAT, and different combinations of these audio features were tested and compared. The researchers in [7] used a cascaded architecture consisting of a Convolutional Neural Network (CNN) followed by recurrent Long Short-Term Memory (LSTM) layers. The model was trained on RECOLA, a French speech emotion corpus. The CNN learned the audio features from raw utterances, which replaced traditional hand-engineered feature extraction – a process dubbed "end-to-end speech emotion recognition."

The authors in [8] also used a cascaded system by studying various combinations of an SVR model and a Bidirectional LSTM-Deep RNN (BLSTM-DRNN). One implementation, dubbed "dependent training," used the first model's prediction output to be fed to the second model's input, along with the other audio features. On the other hand, the "independent training" involved the addition of Gaussian white noise to the data and then using that data to train the first model in order to modify the true labels and create pseudo predictions. These pseudo predictions were then passed as features into the second model along with the other audio features.

The work in [9] used six SVM classifiers for implementing an OVA binary classification strategy. The algorithms were trained on the LDC speech emotion corpus. The radial basis function (RBF) kernel was used in all six SVM models, and the output of each SVM was a confidence score indicating how much the input matched to the emotion class the SVM was trained on. The final prediction came from the SVM model that gave the highest confidence. The performance of the final model was compared with naïve human coders. In [10], the researchers encouraged children with ASD to participate in a virtual game scenario. Through the game, the researchers hoped to teach the participants how to use facial expressions, body language, and vocal tone to express different emotions. The results of the study suggested that there was an improvement in the emotion recognition and socialization skills of the participating children.

The authors of this paper also showed an initial SVM approach for speech-emotion recognition, as seen in [11]. The initial work consisted of evaluating the speech-emotion recognition using 26 MFCCs and zero-crossing rate features to the classifier. The model performed at a peak test accuracy of 77% with clean audio samples and 64% test accuracy when background noise was added to the audio samples. The samples were from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) speech corpus.

III. Audio Processing

A set of custom features was used for this work. The features were selected from a pool of unconventional audio features through trial-and-error. The result was a collection of 36 low-level descriptors (LLDs).

- *MFCCs*: The first 26 MFCCs were extracted for each audio frame using the Hidden Markov Model Toolkit (HTK) implementation [12].
- *Spectral contrast*: It represents the relative spectral distribution [13]. Seven spectral contrast values were extracted from each audio frame.
- *Polynomial coefficients*: Coefficients of fitting an n^{th}-order polynomial to the columns of a spectrogram. Two polynomial coefficients were extracted from each audio frame for a polynomial of order one [14].
- *RMS energy*: The root-mean-square energy of each audio frame. One RMS energy was extracted for each frame.

Librosa, a Python library for music and audio analysis, was used to extract the above mentioned low-level descriptors from speech samples [15]. In this work, a sampling rate of 16 kHz was selected to process each audio file. Each audio frame was 32 ms (512 samples) long and had a 16 ms (256 samples) step size. For performing FFT, 512 samples were considered per audio frame.

Functionals refer to the functions that are applied to a vector. Some of the popular functionals include maximum, minimum, mean, median, standard deviation, among others. Since an utterance is divided into multiple audio frame during processing, the only way to get the feature values for the entire audio signal is to apply functionals on the low-level descriptors.

The mean and the standard deviation were used for the MFCCs. For the spectral contrast, polynomial coefficients, and RMS energy, only the mean was used. The total of 62 audio features in the custom feature set are summarized in Table I.

TABLE I. AUDIO FEATURES IN THE CUSTOM FEATURE SET

Low-Level Descriptors	*Functionals*	*Audio Features*
26 MFCCs	Mean, standard deviation	52
7 Spectral Contrast	Mean	7
2 Polynomial Coefficients	Mean	2
1 RMS Energy	Mean	1

IV. DATASET

Three datasets were considered for the study. The RAVDESS, TESS, and the CREMA-D datasets provide a variance of audio samples to develop and analyze the models. The audio processing of the samples was performed on these three datasets. This section describes the details and considerations of each dataset and the variances of the samples.

A. RAVDESS Dataset

In the year 2018, researchers from Ryerson University's SMART Lab released the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS) [16]. It is a simulated, multi-modal dataset that contains both video and audio data. The WAV audio format was used to record the speech data, with a sampling rate of 48 kHz and a bit depth of 16 bits per sample. A total of 24 actors of age ranging from 21 to 33 years are part of the dataset, where it is split into equal samples of male and female voices. Two lexically similar sentences were recorded by each actor in eight different affective states. There are eight emotion labels to consider: neutral, calm, happy, sad, anger, fear, disgust, and surprise. Out of the eight emotions, seven were recorded twice per sentence – typical and stronger intensities. This project work considers the emotions from the previous work found in [17][18][19]. This gives a total of 1,440 recording samples, with *24 actors x two sentences x eight emotions x two repetitions x two emotional intensities* except for the neutral emotion. Therefore, a total of 192 data samples were created for seven emotions excluding neutral. The neutral class contributed with only 96 utterances. For balancing the dataset, data resampling was used on the neutral samples [20].

B. TESS Dataset

The second speech corpus used for this work was the Toronto Emotional Speech Set (TESS). This simulated dataset was created in 2010 by researchers from the University of Toronto's Psychology Department [21]. It contains recordings from two female actors. During the recordings, the two actors were 26 and 64 years old. A total of 2,800 sentences were recorded in seven different emotions – anger, disgust, fear, happy, surprise, sad, and neutral. TESS is a balanced dataset with 400 data samples per emotion class. However, the utterances recorded by the actors are lexically similar. For each of the seven emotions, the phrase "*Say the word___"* was recorded by each actor 200 times, with a different target word each time [22]. All sentences were recorded at 16-bits per sample, at a sampling rate of 24,414 Hz, and saved in WAV format. The output labels were obtained from the file names.

C. CREMA-D Dataset

The third and final corpus used for this work was the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). Researchers from three different universities worked together to create this speech emotion dataset [23]. Participants had ages ranging from 20 to 74 years old. There were 48 male actors and 43 female actors, a total of 91 participants. And even though they had different ethnic backgrounds, they were all English speakers. CREMA-D, like RAVDESS, is a multimodal dataset. The speech recordings were performed at 16-bits per sample, at a sampling rate of 16 kHz, and saved in the WAV audio format. Twelve emotionally neutral sentences were recorded in six different affective states: happy, sad, neutral, anger, fear, and disgust. Except for the neutral emotion, the five emotions were captured at four different intensity levels. Each emotion class has 1,271 data samples, except for the neutral class, with 1,087 data samples. The dataset was balanced using resampling [24]. Table II gives a short summary of the three datasets used for this work.

TABLE II. SUMMARY OF THE DATASETS

Corpus	*Emotions*	*Samples per Class*	*Total Balanced*
RAVDESS	8 (calm excluded)	192	1,248
TESS	7	400	2,800
CREMA-D	6 (surprise missing)	1,271	7,442

D. Noise Integration

All speech recordings in the above-mentioned corpora were conducted in a quiet, "noiseless" environment. Besides the static coming from the recording equipment and the soft voice echoes, no other sounds can be heard in the background of the clean speech recordings. Being in a "noiseless" environment in practical scenarios can be difficult as there can be multiple noise sources in the background. Therefore, it is essential to integrate background noises into the samples when constructing the final speech emotion recognition model.

Three noise samples were handpicked for the purpose of this study. All three audio files were downloaded from the same site found in [25]. The first noise sample is called "*Small Crowd*," and it was recorded in a playground environment. The second noise recording, "*Shopping Mall Ambiance*," was taken at mall. The final noise sample, titled "*Streets*," was collected at a sidewalk, to capture the sound of cars driving by. The playground noise sample and the street file sample were listed under an attribution 3.0 license, while the shopping mall file sample was listed under a public domain license. The reason for selecting these specific noise samples was to include some of the general properties of common background noises in the audio analysis.

Three signal-to-noise ratios (SNRs) were selected: 0 dB, 5 dB, and 10 dB. Noise samples were added to the clean speech data using these SNRs. The noise addition was performed using

MATLAB by selecting an SNR value randomly out of the three values, and then selecting a noise sample out of the three noise files. The noises were then added to each data sample. To prevent the models from learning the noise features, two measures were taken. The first measure refers to the way the noise samples were added to the clean speech file. Each original noise recording was longer than fifteen seconds, with a duration of 16 seconds for "*Shopping Mall Ambiance*," 19 seconds for "*Small Crowd*," and 49 seconds for "*Streets*". The clean speech samples for all three datasets were roughly three to five seconds long. So, the algorithm randomly selected sections of the noise audio of the clean speech samples' size and added both together to create the noise-added samples. The second measure involves generating only one noise-added sample per clean speech sample. This resulted in the final dataset containing equal amounts of clean speech and noise-added samples. Table III summarizes the details of the final dataset.

TABLE III. DATASETS WITH NOISE INTEGRATION

Corpus	*Data Samples*
RAVDESS (R)	1,248
TESS (T)	2,800
CREMA-D (CD)	7,442
Combined (R+T+CD)	11,490
Combined & Noise-Added (R+T+CD)	11,490
Combined + Noise-Added	22,980
Final Balanced	26,082

V. MACHINE AND ENSEMBLE LEARNING MODELS

A. SVM Model

The development and training of the SVM model consisted of feeding the algorithm extracted features from the audio samples. The SVM model was calibrated to have the best parameters by using the radial basis function (RBF) kernel, with C=10.0 and γ=0.01. The reason for the SVM evaluation was due to the compelling results presented in the literature, which is evident from the performance of the model with the collected datasets and implementation of the background noises.

B. MLP Model

Generally, the number of input-layer neurons is equal to the number of input features, and the number of output-layer neurons is equal to the number of classes in the dataset. When it comes to selecting the size of the hidden layer(s), there is no specific rule that needs to be followed. There are certain conventions among researchers that are listed in [26]. To design the MLP architecture for this work, different number of neurons were tested for each hidden layer, such as 10, 50, 100, 200, and 500. The final model has 62 units in the input layer, which is exactly the number of input features in the custom feature set. The first hidden layer has 105 units, which is around 170% of the input units. The second hidden layer has 62 units, which is equal to the number of input units. Lastly, due to the seven emotion classes used, the output layer consists of seven units.

The categorical cross-entropy loss was used as the loss function, and Adam was used as the optimizer. The Rectified Linear Unit (ReLU) activation was used in the hidden layers. The Softmax activation function gave the prediction scores for each class in the output layer. While training the model, the learning rate was changed with the help of a learning rate scheduler. An inverse time decay function was chosen as the learning rate schedule, with an initial rate of 0.01, decay steps of 1000, and a decay rate of 80%. A batch of size of sixteen was used for the training, validation, and test splits. For training the model, 50 epochs were used along with a 30% dropout between the two hidden layers, and a 10% dropout between the second hidden layer and the output layer.

C. RNN Model

For the RNN model, the time steps were the individual audio frames. The low-level descriptors which were calculated for every audio frame served as the features for the RNN. Using the custom feature set resulted in 36 low-level descriptors per frame, and all 36 of these values were used as audio features for each time step. Since an RNN can only process sequential data, the low-level descriptors (features) are extracted per audio frame (time step) and fed to the network in a sequential manner. This way, for a given data sample, the RNN learns to recognize the pattern of changing feature values with each successive audio frame. With a sampling rate of 16 kHz and a frame length of 512 samples (32 ms), the result was around 150 audio frames per utterance. The input layer has 36 units, corresponding to the 36 low-level descriptors of the custom feature set. The first LSTM layer has 36 cells, equal to the number of low-level descriptors, while the second LSTM layer has twelve cells, which is one-third of the low-level descriptors. The output layer has seven units.

Adam was used as the optimizer and categorical cross-entropy was used as the loss function. Following the standard LSTM cell construction, the hyperbolic tangent (*tanh*) function and the sigmoid (σ) function were used in the LSTM cells as the normal and recurrent activation, respectively. Softmax was used as the activation function for the output units. The inverse time decay function was the learning rate scheduler of choice, with the same parameter values from the MLP model. Data samples from the training, test, and validation partitions were grouped into batches of 16, and 50 epochs were used for training the RNN. A dropout of 30% was set between the two LSTM layers and a 30% dropout was placed between the second LSTM layer and the output. A recurrent dropout of 20% was set for the LSTM cells in the first layer.

D. Ensemble Learning

Voting was applied to combine the SVM, MLP, and RNN models. In voting, the input data is fed to all the classifiers, and predictions are made separately. The class which was predicted the most is selected as the ensemble model's final prediction. During training, the means and standard deviations of all the features were saved along with the trained models. The first step after recording an utterance is to compute the low-level descriptors from every single audio frame. After that, the system applies the necessary functionals on the low-level descriptors to create the audio features.

All audio features are standardized using the means and standard deviations of the features from training. The scaled audio features are sent to the inputs of the SVM and MLP, and the scaled low-level descriptors are sent to the RNN's input. Once all three models have made the predictions, the class that appears the most among the three predictions is declared as the input audio's emotion label. When all three predictions are different, the output of the MLP is chosen as the final prediction since MLP had the highest value for the precision metric. When the Python program is running, it records three seconds of audio continuously and feeds it to the speech emotion recognition system. The program can be stopped anytime by pressing a keyboard interrupt. The pseudo-code for the speech emotion recognition system is in Algorithm 1.

Algorithm 1: Ensemble Learning Execution Code

```
Run a loop indefinitely until interrupt:
    Record microphone audio for three seconds
    Extract low-level descriptors from audio
    Apply functionals to low-level descriptors
    Standardize LLDs with mean and SD from training
    Standardize output of functionals with mean and SD from training
    Feed LLDs to RNN
    Feed functionals of LLDs to SVM and MLP
    Get predictions from all three models
    If no mode available:
        Print emoji corresponding to MLP's predicted label
    If mode exists:
        Print emoji corresponding to the mode of all three model outputs
    Repeat till keyboard interrupt
```

VI. Results & Conclusions

A. SVM Results

The learning curves of this model show that the validation accuracy closely follows the training accuracy in Fig. 2 (*left*). The values selected for the γ and C parameters made sure there was no overfitting during the training process. The test split was equal to 10% of the entire corpus; it contained a total of 2,604 data samples. After stratifying the test set for seven emotion classes, the result was 2,604/7 = 372 data samples per emotion class. From the confusion matrix of Fig. 2 (*right*), surprise appears to be the most accurately detected emotion, followed by anger, sadness, and neutral.

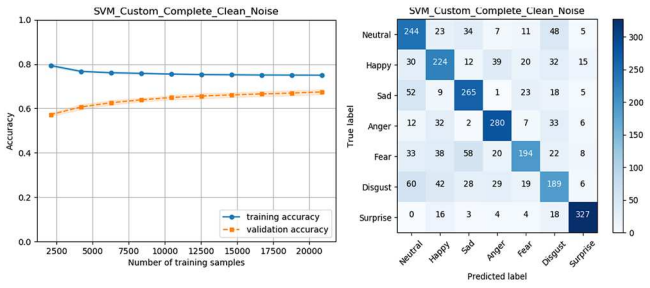

Fig. 2. SVM accuracy training (*blue*) and validation (*orange*) curves (left) and the confusion matrix (*right*) of the resulting classification performance of the testing splits over the seven discrete emotions.

TABLE IV. SVM Performance Comparision Over All Datasets

Corpus	*Train%*	*Valid %*	*Test %*	*Precision*	*Recall*
RAVDESS Clean	99.0 %	85.8 %	80.6 %	81.8 %	80.6 %
RAVDESS Clean + Noise	87.0 %	61.6 %	58.2 %	58.9 %	58.2 %
TESS Clean	100 %	99.6%	100 %	100 %	100 %
TESS Clean + Noise	99.0 %	98.6 %	97.7 %	97.7 %	97.7 %
CREMA-D Clean	86.0 %	54.9 %	57.6 %	57.7 %	57.6 %
CREMA-D Clean + Noise	79.0 %	51.1 %	52.1 %	51.9 %	52.1 %
Complete Clean	81.0 %	71.9 %	73.0 %	74.3 %	73.0 %
Complete Clean + Noise	**75.0 %**	**65.2 %**	**66.1 %**	**66.4 %**	**66.1 %**

B. MLP Results

From Fig. 3, it is apparent that the gaps between the accuracy curves and the loss curves are very small. This was not the case during the first round of experiments as no regularization was used at that time. The huge overfitting issue was exposed by the resulting curves and classification scores. After using some dropout between the layers, the gap between the training and the validation accuracy was reduced. Another problem was the fixed learning rate, which prevented the loss function from converging to its minimum value during the final phase of the training process. The rising loss curves were a clear indication of this problem. With the help of a learning rate scheduler, the learning rate was gradually decreased throughout the training, which kept reducing the training and validation losses. Fig. 4 (*left*) show the confusion matrix and performance of precision and misclassification values for the MLP test evaluation. Again, the top three most accurately classified emotions were surprise, anger, and sadness.

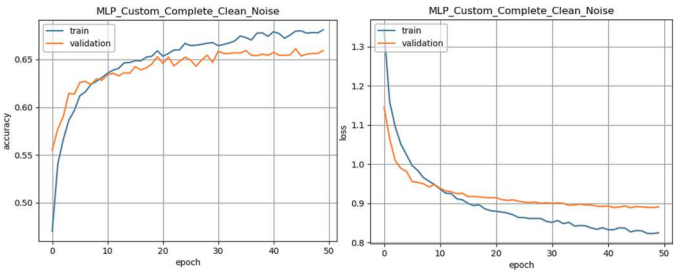

Fig. 3. MLP accuracy (*left*) and loss (*right*) curves for training (*blue*) and validation (*orange*), progression over the number of epochs.

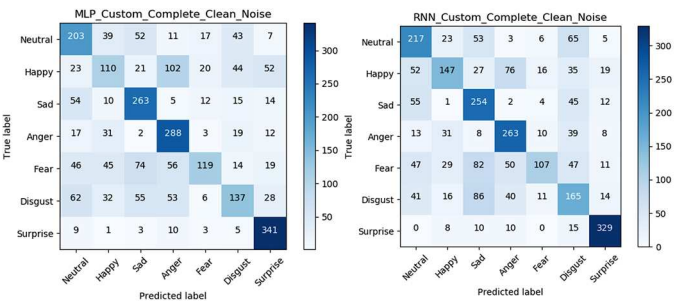

Fig. 4. MLP (*left*) and RNN (*right*) confusion matrices of the resulting classification performance of the testing splits over the seven discrete emotions.

The test accuracy of the MLP was about one percent less than that of the SVM model as seen in the last row from Table V. However, the MLP achieved a higher precision score. As the validation loss became almost steady after the 50th epoch, the training was stopped at that point. It roughly took the same amount of time to train the MLP and SVM. An interesting observation of the performance of the MLP model is that it can handle the noise background to the audio samples with greater efficiency than the clean audio samples in one of the three datasets.

TABLE V. MLP PERFORMANCE COMPARISION OVER ALL DATASETS

Corpus	*Train%*	*Valid %*	*Test %*	*Precision*	*Recall*
RAVDESS Clean	94.5 %	82.8 %	77.6 %	77.9 %	76.1 %
RAVDESS Clean + Noise	80.6 %	64.2 %	56.7 %	66.3 %	50.0 %
TESS Clean	99.9 %	100.0 %	100.0 %	100.0 %	100.0 %
TESS Clean + Noise	99.5 %	98.0 %	98.8 %	98.8 %	98.8 %
CREMA-D Clean	72.7 %	54.5 %	54.7 %	61.5 %	44.5 %
CREMA-D Clean + Noise	60.7 %	51.9 %	54.5 %	65.7 %	36.2 %
Complete Clean	75.9 %	70.8 %	69.5 %	80.5 %	60.0 %
Complete Clean + Noise	**68.1 %**	**65.9 %**	**65.7 %**	**83.3 %**	**50.3 %**

C. RNN Results

Just like the MLP model, the training of the RNN model was halted after 50 epochs, as the validation loss stopped decreasing after that, as shown in Fig. 5, with accuracy curves (*left*) and loss curves (*right*). Using too many epochs during training can cause a serious overfitting problem as the model will become too familiar with the training samples and struggle to accurately predict the labels of unseen data. The RNN model took the longest time to train among all three models.

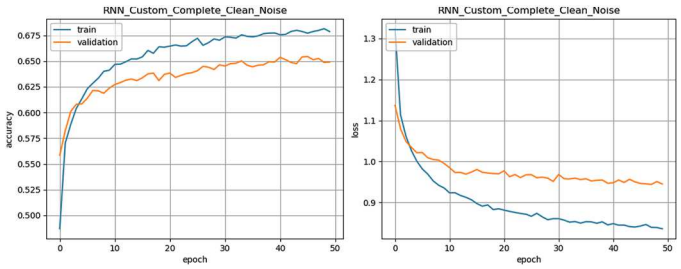

Fig. 5. RNN accuracy (*left*) and loss (*right*) curves for training (*blue*) and validation (*orange*) progression over the number of epochs.

The average precision score of the RNN model was higher than that of the SVM model, as noted in Table VI. The RNN slightly lags behind the MLP model when it comes to classification scores and precision value. However, the RNN model beats the MLP in the average recall score. Like the SVM and MLP model, the top three most accurately predicted emotions for the RNN model were surprise, anger, and sadness, in the written order. The confusion matrix for the RNN is shown in Fig. 4 (*right*).

TABLE VI. RNN PERFORMANCE COMPARISION OVER ALL DATASETS

Corpus	*Train%*	*Valid %*	*Test %*	*Precision*	*Recall*
RAVDESS Clean	94.6 %	76.1 %	61.2 %	61.2 %	59.0 %
RAVDESS Clean + Noise	85.0 %	59.3 %	51.5 %	55.9 %	49.6 %
TESS Clean	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %
TESS Clean + Noise	99.8 %	99.3 %	99.6 %	99.6 %	99.6 %
CREMA-D Clean	67.7 %	49.0 %	55.0 %	58.3 %	47.1 %
CREMA-D Clean + Noise	59.9 %	51.2 %	51.0 %	58.7 %	40.5 %
Complete Clean	74.3 %	66.5 %	64.5 %	72.1 %	57.3 %
Complete Clean + Noise	**67.9 %**	**64.9 %**	**63.7 %**	**75.4 %**	**53.7 %**

D. Ensemble Learning Results

Table VII compares the ensemble learning model with the individual classifier models trained on the complete dataset with the noise background. The ensemble model can be seen to have achieved 0.8 % higher accuracy than the MLP model, which performed the best among the individual models. This improvement of 0.8 % is equivalent having twenty more accurately classified samples out of the 2,607 test samples.

TABLE VII. TEST ACCURACIES OF THE INDIVIDUAL MODELS AND ENSEMBLE LEARNING OUTCOME

Model	*Test Accuracy*
SVM	66.1%
MLP	65.7%
RNN	63.7%
Ensemble	66.5%

E. Conclusions

It can be seen from the results that the MLP model had the highest precision score, while the SVM model showed the lowest precision score. The SVM, MLP, and RNN models performed extremely well on the TESS corpus. TESS is made up of recordings from only two female participants who recorded 200 similar-sounding sentences in each of the seven affective states, with no change in intensity. This lack of variation did not pose any problems for the models during prediction as they quickly picked up the data samples' features. All three individual models showed the lowest performance on the CREMA-D dataset. Unlike TESS, this dataset had contributions from a lot of participants, with 48 male actors and 43 female actors. Furthermore, each recording in CREMA-D was performed in four different emotional intensities. The models clearly struggled to learn all these variations in the data.

The other observation regarding the performances of the models is that it reflects the confusion in detecting similar emotions in the same manner as humans. The correlation of anger to disgust, or sad to fear can present a problem in accurately classifying emotions. Other issues can arise when dealing with more complex emotions, those which humans can easily understand, like when someone is being sarcastic or upset

when smiling. Features of the voice will need to be examined further to find out which features are essential for which emotion groups and which are inefficient in categorizing these emotions. Emotions are not exclusively detected based on speech recognition, but they also manifest in facial and body language, something that will need to be included to increase the performance of the classification to clearly segregate human emotions through machine learning.

ACKNOWLEDGMENT

We would like to acknowledge and thank Dr. Vishu Viswanathan from the Ingram School of Engineering for his valuable advice in this project. Also, we want to thank Dr. Maria Resendiz and Anna Stewart from the Communication Disorders Department at Texas State University for their input and advice in this study.

REFERENCES

[1] N. H. Frijda, *The Emotions*, Cambridge, England, UK: CUP, 1986.

[2] O. Korn, L. Stamm, and G. Moeckel, "Designing authentic emotions for non-human characters. A study evaluating virtual affective behavior," *Designing Interactive Systems*, pp. 477-487, Jun 2017.

[3] A. Mehrabian, *Silent Messages*, Belmont, CA, USA: Wadsworth Pub. Co, 1971.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion", *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, Feb 1971.

[5] N. Kurpukdee, S. Kasuriya, V. Chunwijitra, C. Wutiwiwatchai and P. Lamsrichan, "A study of support vector machines for emotional speech recognition," *2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, pp. 1-6, 2017.

[6] A. Meftah, Y. Alotaibi and S. Selouani, "Emotional speech recognition: A multilingual perspective," *2016 International Conference on Bio-engineering for Smart Technologies (BioSMART)*, pp. 1-4, 2016.

[7] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200-5204, 2016.

[8] J. Han, Z. Zhang, F. Ringeval and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5005-5009, 2017.

[9] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan and W. Heinzelman, "Emotion classification: How does an automated system compare to Naive human coders?," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2274-2278, 2016.

[10] B. Schuller, E. Marchi, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, P. Robinson, I. Davies, T. Baltrusaitis and M. Mahmoud, "Recent developments and results of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions," *IDGEI*, 2015 (No pagination provided).

[11] R. Matin and D. Valles, "A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions," *2020 Intermountain Engineering, Technology and Computing (IETC)*, Orem, UT, USA, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249147.

[12] "What is HTK?" [Online] Available: http://htk.eng.cam.ac.uk/ [Accessed: 18-Feb-2021].

[13] D. Jiang, L. Lu, H. Zhang, J. Tao and L. Cai, "Music type classification by spectral contrast feature," *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 113-116, 2002.

[14] O. Agcaoglu, B. Santhanam and M. Hayat, "Improved spectrograms using the discrete Fractional Fourier transform," *IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, pp. 80-85, 2013.

[15] "librosa," [Online] Available: https://librosa.org/doc/latest/index.html. [Accessed: 25-Jan-2021].

[16] "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," [Online] Available: https://zenodo.org/record/1188976. [Accessed: 23-Feb-2021].

[17] M. I. Ul Haque and D. Valles, "Facial Expression Recognition Using DCNN and Development of an iOS App for Children with ASD to Enhance Communication Abilities," *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York City, NY, USA, 2019, pp. 0476-0482. doi: 10.1109/UEMCON47517.2019.8993051.

[18] M. I. Haque, D. Valles, "Facial Expression Recognition from Different Angles Using DCNN for Autistic Children to Recognize Emotional Patterns," *The 5th Annual Conference on Computational Science & Computational Intelligence – Symposium on Signal & Image Processing, Computer Vision & Pattern Recognition (CSCI-ISPC'18)*, Las Vegas, NV, 2018, pp. 446-449, doi:10.1109/CSCI46756.2018.00090.

[19] M. I. Haque, D. Valles, "A Facial Expression Recognition Approach using DCNN for Autistic Children to Identify Emotions," *The 9th IEEE Annual Information Technology, Electronics & Mobile Communication Conference (IEMCON'18)*, Vancouver, Canada, 2018, pp. 546-551, doi:10.1109/IEMCON.2018.8614802.

[20] 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, pp. 1-35, 2018.

[21] Pichora-Fuller, M. Kathleen and Dupuis, Kate, "Toronto emotional speech set (TESS)," *Scholars Portal Dataverse*, doi: 10.5683/SP2/E8H2MF.

[22] "Toronto emotional speech set (TESS) A dataset for training emotion (7 cardinal emotions) classification in audio," [Online] Available: https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess. [Accessed: 22-Feb-2021].

[23] "CREMA-D (Crowd-sourced Emotional Mutimodal Actors Dataset)," [Online] Available: github.com/CheyneyComputerScience/CREMA-D. [Accessed: 28-Jan-2021].

[24] C. Houwei, 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, pp. 377-390, Jan 2014.

[25] "SoundBible.com," [Online] Available: http://soundbible.com/. [Accessed: 7-Mar-2021].

[26] "The number of hidden layers," [Online] Available: https://www.heatonresearch.com/2017/06/01/hidden-layers.html. [Accessed: 12-Mar-2021].