I. DATASET

In this work, we use the IEMOCAP released in 2008 by researchers at the University of Southern California (USC). It contains five recorded sessions of conversations from ten speakers and amounts to nearly 12 hours of audio-visual information along with transcriptions. It is annotated with eight categorical emotion labels, namely, anger, happiness, sadness, neutral, surprise, fear, frustration and excited. It also contains dimensional labels such as values of the activation and valence from 1 to 5; however, they are not used in this work.

The dataset is already split into multiple utterances for each session and we further split each utterance file to obtain wav files for each sentence. This was done using the start timestamp and end timestamp provided for the transcribed sentences.

II. METHODOLOGY

A. Data Pre-processing

A preliminary frequency analysis revealed that the dataset is not balanced. The emotions “fear” and “surprise” were under-represented and use upsampling techniques to alleviate the issue. We then merged examples from “happy” and “excited” classes as “happy” was under-represented and the two emotions closely resemble each other. In addition to that, we discard examples classified as “others”; they corresponded to examples that were labelled ambiguous even for a human.

B. Feature Extraction

1. Pitch

Pitch is important because waveforms produced by our vocal cords change depending on our emotion. Many algorithms for estimating the pitch signal exist. We use the most common method based on autocorrelation of center-clipped frames.

2. Harmonics

In the emotional state of anger or for stressed speech, there are additional excitation signals other than pitch. This additional excitation is apparent in the spectrum as harmonics and cross-harmonics. We calculate harmonics using a median-based filter.

3. Speech Energy

Since the energy of a speech signal can be related to its loudness, we can use it to detect certain emotions. We use standard Root Mean Square Energy (RMSE) to represent speech energy. RMSE is calculated frame by frame and we take both, the average and standard deviation as features.

4. Pause

We use this feature to represent the “silent” portion in the audio signal. This quantity is directly related to our emotions; for instance, we tend to speak very fast when excited (say, angry or happy, resulting in a low Pause value).

C. Machine Learning Models

1. Random Forest (RF):

Random forests are ensemble learners that operate by constructing multiple decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees.

2. Gradient Boosting (XGB):

XGB refers to eXtreme Gradient Boosting. It is an implementation of boosting that supports training the model in a fast and parallelized way. Boosting is another ensemble classifier combining a number of weak learners, typically decision trees. They are trained in a sequential manner, unlike RFs, using forward stagewise additive modeling. During the early iterations, the decision trees learned are simple. As training progresses, the classifier becomes more powerful because it is made to focus on the instances where the previous learners made errors. At the end of training, the final prediction is a weighted linear combination of the output from the individual learner.

3. Support Vector Machines (SVMs):

SVMs are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. An SVM training algorithm essentially builds a non-probabilistic binary linear classifier.

D. Deep Learning Models

1. Multi-Layer Perceptron (MLP):

MLP belongs to a class of feed-forward neural network. It consists of at least three nodes: an input, a hidden and an output layer. All the nodes are interleaved with a non-linear activation function to stabilize the network during training time. Their expressive power increases as we increase the number of hidden layers up to a certain extent. Their non-linear nature allows them to distinguish data that is not linearly separable.

2. Long Short-Term Memory (LSTM):

LSTMs were introduced for long-range context capturing in sequences. Unlike MLP, it has feedback connections that allow it to decide what information is important and what is not. It consists of a gating mechanism and there are three types of gates: input, forget and output.

III. IMPLEMENTATION

- We use librosa, a Python library, to process the audio files and extract features from them.
- We use scikit-learn and xgboost, the machine learning libraries for Python, to implement all the ML classifiers (RF, XGB and SVM) and the MLP.
- We use PyTorch to implement the LSTM classifiers described earlier.
- In order to regularize the hidden space of the LSTM classifiers, we use a shut-off mechanism, called dropout, where a fraction of neurons is not used for final prediction. This is shown to increase the robustness of the network and prevent overfitting.

We randomly split our dataset into a train (80%) and test (20%) set. The same split is used for all the experiments to ensure a fair comparison. Different batch sizes were used for different models.

IV. RESULTS

Performance of LSTM reveals that deep models indeed need a lot of information to learn features as the LSTM classifier trained on eight-dimensional features achieves very low accuracy as compared to the end-to-end trained ARE. However, neither of them is able to beat the lighter E1 model (Ensemble of RF, XGB and MLP) which was trained on the eight-dimensional audio feature vectors. A look at the confusion matrix reveals that detecting “neutral” or distinguishing between “angry”, “happy” and “sad” is the most difficult for the model.