Multi-label Emotion Classification using modified Sequential Deep Learning Architecture for Speech Emotion Recognition.

I. ABSTRACT

In recent years, audio classification has gained significant attention due to its wide range of applications, particularly in the field of Speech Emotion Recognition (SER). This paper presents a novel approach to multi-label audio classification based on emotion using a modified Sequential Deep Learning architecture, applied to a combined dataset of TESS, RAVDESS, SAVEE, and CREMA-D datasets. Our model employs SoftMax to obtain multi-class probabilities, which are then used to calculate evaluation metrics and uncertainties. Additionally, we explore the performance of our model across various emotion categories, demonstrating its robustness and versatility. Our model used Dropout Layers after every Layer in the architecture to decrease the overfitting and increase the accuracy. Comprehensive experiments show that our model achieves state-of-the-art results in multi-label classification, outperforming previous methods. The method also extensively evaluates the model's performance, including metrics such as accuracy, precision, validation accuracy, F1-score, and recall.

***Index Terms*—Audio Classification, SER, Combined Dataset, Multi-Class.**

II. INTRODUCTION

Speech Emotion recognition (SER) [1] is an ever-improving field finding itself place in a wide range of applications, from lie detection and threat assessment to customer care, gaming, marketing, tele-healthcare, education, and sports. It may also lead to direct interaction with the computer or any artificial intelligence verbally especially now that the AI interactions are becoming a daily part of our life. Recent advancements in deep learning models, such as Convolutional Neural Networks and Recurrent Neural Networks, have made significant improvement in the accuracy of SER systems. As we are making further advancements, we pave the way for more advanced applications that leverage the capability to recognize and respond to human emotions. Despite these advancements, there are still many challenges that need to be addressed to further enhance HMI [2]. Proposing effective solutions to improve the integration of SER into real-world applications is crucial. There are various models for emotion recognition tasks but there are still many problems to be addressed [3], for this reason we need to propose an effective solution to increase the Human-Machine Interaction (HMI). In this model we implemented a model [9] that employs Convolution layers for feature extraction, Dense layers for deeper learning, and MaxPooling for dimensionality reduction, each with dropout and L2 regularization to prevent overfitting. The Dropout layers are very important because they decrease the overfitting of the data which occurs due to mixing of data from various sources and unnecessary extra features. Finally, the SoftMax layer outputs probabilities across 5 classes.

III. RELATED WORKS

In the domain of multi-label audio classification, several methodologies have been developed to enhance performance by leveraging deep learning architectures because deep learning architectures and learning techniques are very famous in this domain of SER. Numerous studies have explored multi-label classification in Speech Emotion Recognition, each aiming to address the core challenge of accurately capturing relationships between multiple features.

In a study by Cohn et al. (2009), pioneered the detecting depression from facial actions and vocal prosody [4] using Active Appearance Modeling (AAM) and manual Facial Action Coding System (FACS), highlighting the effectiveness of facial and vocal features in identifying depressive symptoms [4].

Additionally, Cohn and De La Torre (2015) expanded on this approach by integrating multimodal features, achieving significant accuracy improvements [5].

In a study on Speech Emotion Recognition (SER) using deep convolutional neural networks, researchers explored the effectiveness of various models, including the RAVDESS, EMO-DB, and IEMOCAP datasets. The study achieved notable accuracy levels, with the RAVDESS model reaching 67 percent accuracy and the EMO-DB model achieving 92.86 percent [19] [20].

Analysis of Vocal Pattern to Determine Emotions using Machine Learning made a model that utilises CNN and SVM to extract features and the data was split into training and testing and achieved an accuracy of 81 percent and 65 percent respectively. They categorized human emotion by several attributes such as pitch, timbre, loudness, and vocal tone [6].

Maisy Wieman, Andy Sun in Analyzing Vocal Patterns to Determine Emotion: in their framework utilised Support

Vector Machine, Generalized Linear Model, Binary Classification Tree. They Achieved good accuracy mainly using Binary Decision Tree for prediction. Their model contributed to great decrease in the error rate in the training data [7].

Speech Emotion Recognition Based on Multi-feature and Multi-lingual Fusion, authored by Zhang et al., employs models utilizing Low-Level Descriptors (LLD), VGGish, and a fusion of LLD+VGGish features. The study leverages four datasets—CASIA, CHEAVD, LEMOCAP, and SAVEE—achieving an average accuracy of 55.4 percent. The authors highlight the effectiveness of multi-feature and multilingual fusion for emotion recognition, though achieving consistent accuracy across different languages and feature sets presents challenges [19].

Zhao et al,H.Ranganathan,S.Chakraborty used DemoFBV, DemoFV, and CDBN models in the realization of multimodal emotion recognition using Deep Learning Architectures at 90 percent, 84 percent, and 97 percent accuracies, respectively. The authors have further proposed that the fusion of the complementary features outweighs the weight attached to them. This has been based on the argument that the fusion takes into account the complex emotional cues and, hence, significantly improves the classification accuracy across different categories of emotions [1].

Li et al. carry out deep convolutional neural networks operation on the RAVDESS and EMO-DB datasets in their work with speech emotion recognition, realizing accuracies of 67 percent and 92.86 percent, respectively, although compatibility turned out to be low on the IEMOCAP dataset and achieved unsatisfactory accuracy by the current model setting. The authors demonstrate that performances of models differ for each dataset, which can range widely to ensure dependable classification of emotions based on choice of datasets [3].

Inspired by these advancement, we propose the combination of multiple sequential deep learning techniques to achieve a multi-class emotion prediction. Our approach bases dropout layer as the backbone that reduces overfitting and on the other hand Convolution layer analyzes and assigns scores to possible label combinations. This enables the model to differentiate among many classes, thereby improving upon high classification accuracies in complex scenes.

IV. Dataset and Methodology

We have taken 4 datasets here which include- TESS, SAVEE, RAVDESS, CREMA-D commonly used for emotion recognition. From the given four datasets we can extract 7 kinds of emotions which are happiness, sad, disgust, fear, anger, calm and neutral but we only consider five emotions for the final model. We take the emotions 'angry', 'calm', 'happy', 'sad', 'surprise' for further processing. We perform the comparison of audio and visual features in the datasets followed by the deep learning model [10]. We used a combination of convolutional and fully connected layers, dropout for regularization and 12 regularization to manage the overfitting. The output of the final layer gives the probabilities for the 5 emotions taken from the dataset.

A. Dataset

The TESS dataset consists of audio recordings from two actresses, each 26 years and 64 years of age, where they recorded 2800 utterances in the English language. Both were made to say a carrier phrase, "Say the word ....", in sets portraying each of the seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. From these data, we have data from four of the seven emotions which are 'anger', 'happiness', 'sad' and 'surprise' [22].

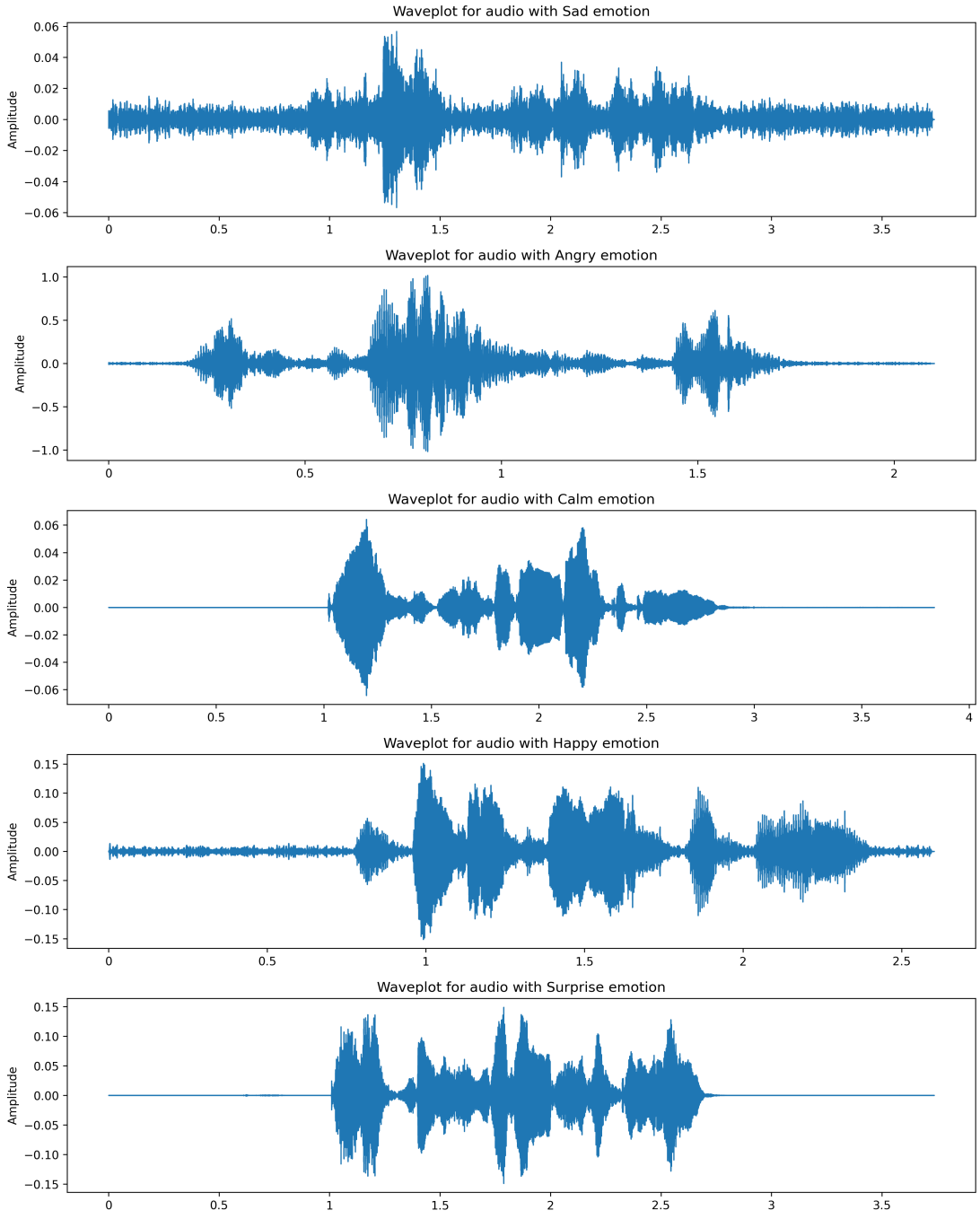
The second data set used in the model is Surrey Audio-Visual Expressed Emotion (SAVEE, which has different approach, provides the voices of 4 male actors recorded from the age of 27 to 31. There were 120 utterances for every speaker totaling 480 utterances in total. There were a total of 7 emotion categories recorded; that were- anger, disgust, fear, happiness, sadness, surprise, and a neutral. Out of these seven emotions, we have derived five emotions for further pre-processing in the model which are- 'angry', 'calm', 'happy', 'sad' and 'surprise'.

The third dataset used for this model is the Ryerson Audio-Visual Database of Emotional Speech And Song (RAVDESS). [9] The dataset is selected because it is readily available. This dataset contains both audio and visual recordings of 12 male and female actors respectively vocalizing statements in English. A total of seven emotions were vocalized of them being calm, happy, sad, angry, fearful, surprise, and disgust. The dataset is divided into speech and song where there are 1440 speech files and 1012 song files. Just like the dataset above we take the five emotions-'angry', 'calm', 'happy', 'sad' and 'surprise only for further model training". [3]

The last dataset used in the model is Crowd Sourced Emotional Multimodal Actors Dataset, or CREMA-D. It has 7442 audio clips from one of the six emotions: Anger, Disgust, Fear, Happy, Neutral, or Sad, and four different emotion levels: Low, Medium, High, and Unspecified. It was expressed by 91 actors out of which 48 are males and 43 are females, where the acting people belong to different ethnic (Asian, African, American e.t.c.), age groups ranging between 20 and 74. Out of these six emotions, 'anger', 'happy', and 'sad' emotions are chosen for further preprocessing [16].

There were total of 7 emotions which were available in all the four datasets together but only five of them were considered because emotions such as neutral, calm and sad, disgust were overlapping with each other causing overfitting of the model. Total of 6373 are left in the end after dropping the unneccesary features

We can visualize the audio dataset through their waveplots:



B. *Methodology*

1) Feature Extraction: Effective feature extraction is vital for the performance of any machine learning model. Selecting relevant features enhances the model's training quality, while irrelevant ones can substantially impede it. For our feature extraction process, we employed the Librosa audio library .Specifically, we utilized five unique spectral representations of the audio as inputs to our deep learning model, which include [3] [8]:

- UMel-frequency Cepstral Coefficients (MFCCs).
- Mel-scaled spectrogram
- Chromagram
- Spectral contrast feature
- Tonnetz [8]

Mel-frequency Cepstral Coefficients (MFCCs) [15] [23]: These are extensively applied in audio and speech processing to represent the sound by taking the short-term power spectrum on the Mel scale. They effectively track changes in timbre variations by applying a Fourier transform to extract the energy spectrum, though they might lack the precision of catching up pitch and harmony.

Mel-scaled Spectrogram: It maps the sound frequency spec-

trum onto the Mel scale so that the output is modeled to mirror human listening. Mel-scaled spectrograms are more suitable for changes in timbre but are more unreliable to denote pitch classes and harmonic details.We can convert a frequency [21] f in Hertz (Hz) to the Mel scale, which is a perceptual scale of pitches that reflects the way humans perceive sound frequencies:

$$\text{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

Chromagram: Chromagrams are good features for pitch and harmonic analysis that often capture harmonic relations missed by MFCCs and Mel-scaled spectrograms. From STFT, it captures pitch classes and harmony at better resolution than the above-mentioned earlier methods.The a way to calculate the Chromagram for a given time t and pitch class k is [3]:

$$C_k(t) = \sum_{\omega \in P_k} |X(\omega, t)|$$

Spectral Contrast: It is a rich feature regarding spectral analysis because it can pick out peaks and valleys in the sound spectrum. It is very useful in the classification of music genres, often outperforming the Mel-scale techniques, while it captures more subtle spectral differences.The way to calculate spectral contrast is : Tonnetz (Tonal Centroid Space):

$$\mathrm{s(t,\, b)} = \log(\mathrm{P_a(t)}) - \log(\mathrm{V_a(t)})$$

It describes harmonies and their pitch dependencies in terms of a six-dimensional pitch space to which tonal centroids are projected. This technique, proposed by Harte et al., places more harmonically related pitches closer in space, which can help with refined sound analysis for applications such as music classification and emotion recognition.

*2) **Model Architecture**:* The first layer of the model contains a Convolutional Layer which consists of two one-dimensional convolutional layers. The first contains 265 filters, and the second one contains 128 filters. They help in extracting important features from the input signal and to further minimize overfitting we applied L2 regularization which adds penalty for larger weights. We also use Dropout Layers which work in manner of randomly "dropping out" (setting to zero) fraction of neurons at each training step which means those neurons temporarily taken out of network. After initial Convolutional layers there's Dropout layer with rate 0.3 which zeros out 30 percent of neurons [6] [18].

$$Z = RELU(W * X + b)$$

The second layer in the model is the Dense Layer[14], in which the main idea behind dense layers is that they allow each neuron to "see" all of the outputs from the previous layer, making them an indispensable part of learning complex patterns. They assist the model in extracting those higher-level abstract features built over the raw features extracted by the convolutional and pooling layers. Dense layer with 128 units Dropout layer dropout rate set to 0.5 Dense layer 64 units Dropout layer dropout rate set to 0.5.

$$Z[i] = max(X[i : i + poolsize])$$

Then we add some more Convolution and Pooling Layers [2]. Here the Convolution Layer further refines features, and the Pooling Layer reduces the dimensionality yet keeping the critical feature positions. We prepare this for output by adding a Flatten Layer. This flattening layer serves as a connecting junction between convolution layers and dense layers. It then transforms multidimensional outputs from convolutional and pooling layers into one-dimensional vectors to be further processed with dense layers for the final steps of classification.

Lastly, we add a final learning dense layer, followed by a dropout layer [13] using a dropout rate of 0.4 in order not to overfit the model. This is then followed by a SoftMax layer as the activation function since multi-class classification problems require outputting a probability distribution across classes for what we target. This SoftMax layer computes the probabilities required for a five-class classification problem. We can visualize the model through the diagram:

$$SoftMax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{5} e^{z_j}}$$

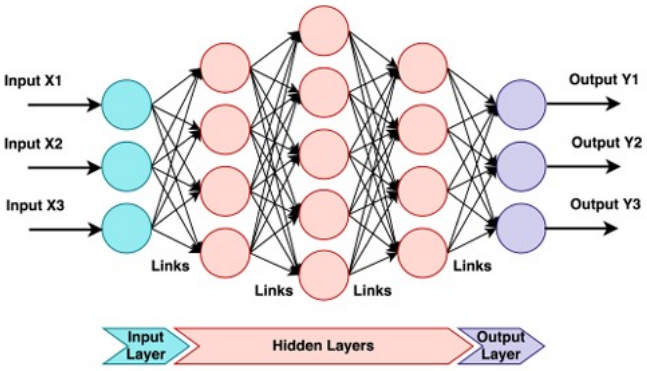


Fig. 1. this is a generic layer wise deep neural network architecture

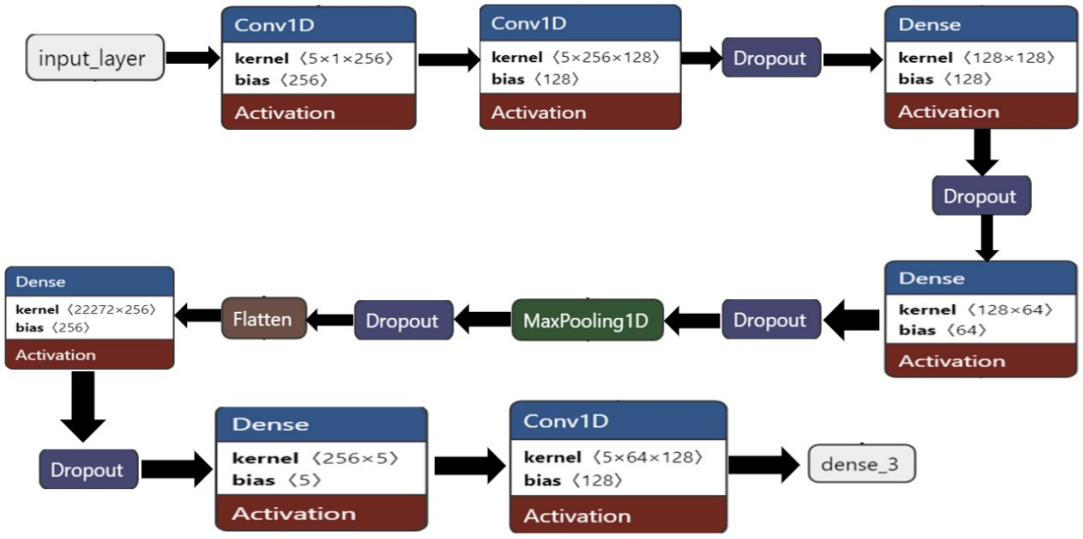


Fig. 2. Pipeline for Model Architecture

V. Results

The multilabel classification of the emotions resulted from our modified Sequential Deep Learning Architecture from the TESS, SAVEE, RAVDESS, and CREMA-D datasets, showing the model's efficiency in emotion prediction. We take a classification report, which contains precision, recall, and f1-score to evaluate the performance of the model. We also calculated the accuracy and the loss functions which are the most important part of the evaluation metrics. We also make use of an ROC Curve that provides a graphical description of the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different thresholds of classification.

*1) **Equations**:* The formula for f1 score is given by:

$$\textbf{F1 Score} = 2 * \frac{Recal * Precision}{Recall + Precision}$$

The formula for precision is given by:

$$\textbf{Precision} = \frac{TruePositives}{TruePositive + FalsePositive}$$

The formula for accuracy is given by:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FN + TN + FN}$$

The formula for Loss Function is:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}\cdot\log(\hat{y}_{ij}) + (1-y_{ij})\cdot\log(1-\hat{y}_{ij}))$$

A. *metrics and graphs*

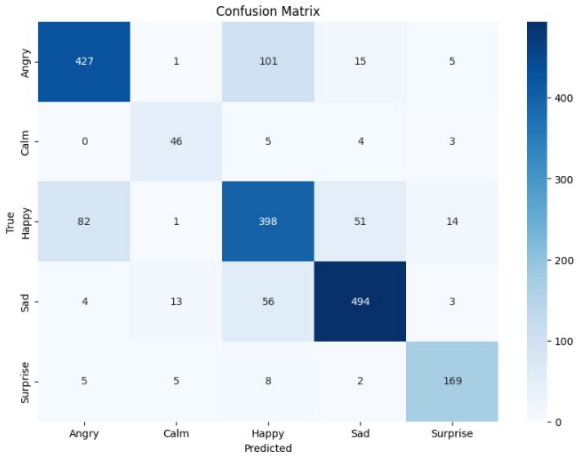


Fig. 3. Confusion Matrix

A confusion matrix is a tabular representation to access the performance of a classification model. It gives a graphical display of the actual against the predicted classifications enabling the possibilities to be known as to how well the model is working. The confusion matrix is really helpful with binary and multi-class classification problems. The confusion matrix shows the model got "Angry" correct 427 times but misclassified as "Happy" 101 times. "Calm" was properly classified 46 times, very few errors. For "Happy", 398 were correct, though 82 were misclassified as "Angry" and 51 as "Sad". "Sad" was highly accurate with 494 correct predictions, "Surprise" had 169 correct predictions with minimal misclassification. Overall, the model goes well, but somehow gets confused between similar emotions like "Happy" and "Sad."

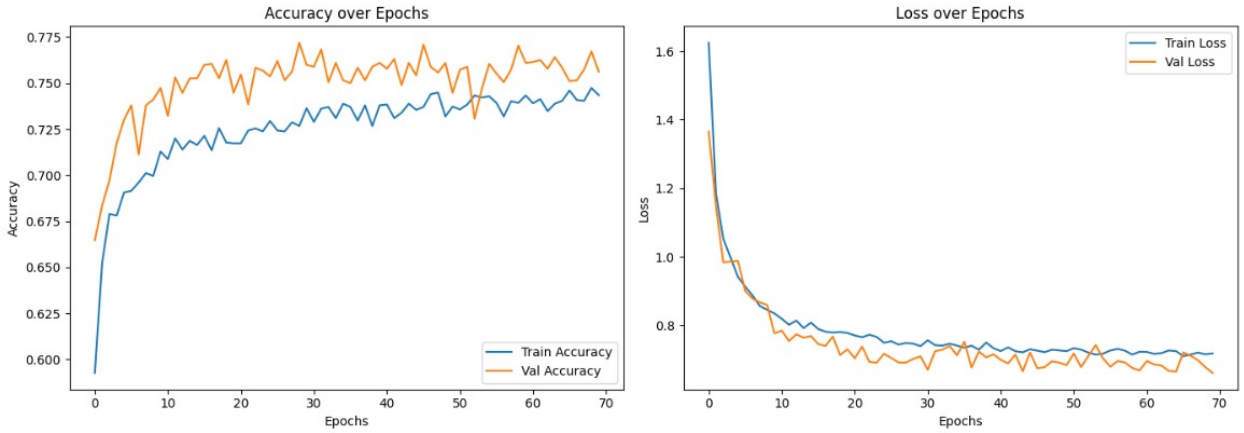


Fig. 4. Track accuracy and loss over epochs for both training and validation sets.

The classification report below generally tends to show fairly good overall performance with a macro average precision, recall, and F1-score of around 0.79. The model performs well for ”Angry,” ”Sad,” and ”Surprise” but fails for ”Calm” and ”Happy,” mainly concerning the precision. Improving performance would require some data augmentation, balancing the dataset, hyperparameter tuning, and possibly more complex models. Feature engineering with domain knowledge could also enhance accuracy for harder classes like ”Calm” and ”Happy.”

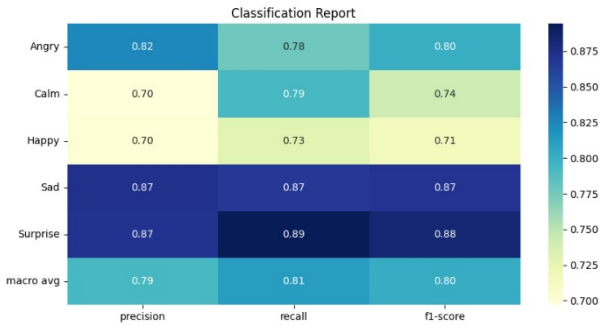


Fig. 5. Classification Report

From the ROC Curve below we get to know that- The curves are in general hugging the upper left corner, which is an indication of very good performance. This is to mean that the classifier is able to establish good discrimination between the positive classes and negative classes. Class Calm and Surprise: These two classes have the highest AUC-ROC values at 0.99, a good indication that they performed well as far as differentiating between the positive and negative instances are concerned. Class Angry and Sad: In general, scores are lower for these classes but still very good (0.95 and 0.97, respectively). Class Happy: This class has the lowest AUC-ROC score of 0.90, which indicates relatively lower performance than that of the other classes.

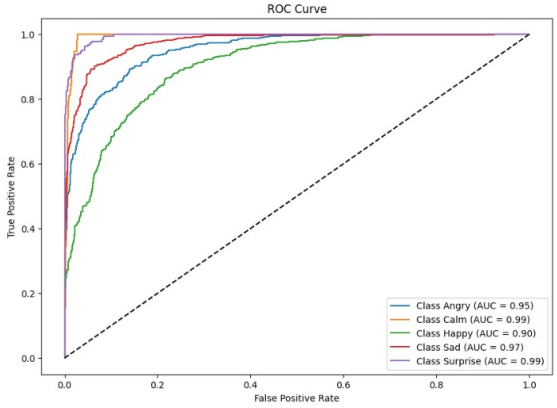


Fig. 6. ROC Curve of all Emotions

The histogram of confidence suggests an overall well-working model that concentrates most predictions closer to 1.0, thus reflecting a high confidence in the generated outputs. A few low-confidence predictions might indicate areas of potential improvement - either at particularly difficult data points or edge cases. Calibration can also potentially increase the correspondence between confidence scores and actual probabilities and thereby improve the interpretability of the model, making it better. [17] Again, the ultimate question is how well the model performs in the domain-specific context, as different applications may demand different confidence levels. On the whole, despite its high confidence, its low-confidence cases must be further refined and calibrated to really make it reliable.

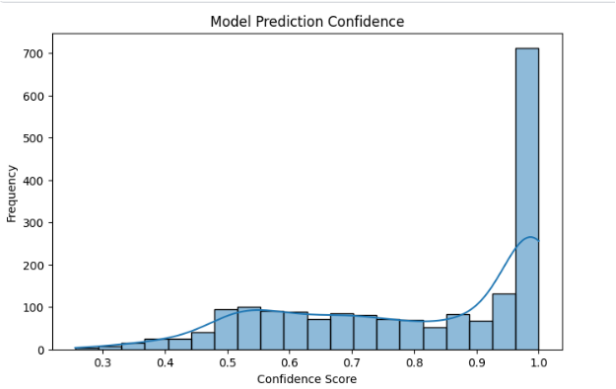


Fig. 7. ROC Curve of all Emotions

The table below shows the classification result for the prediction of emotions: Angry, Calm, Happy, Sad, and Surprise. The model does well for the *Sad* and *Surprise* emotions as the support counts are high with steady F1-scores of 0.87 and 0.88, respectively. The model is having a bit of trouble in the Calm category since it has fewer support counts (58 instances), consequently lowering the performance of the other emotions as well.

Emotion	precision	recall	f1-score	support
Angry	0.82	0.78	0.80	549
Calm	0.70	0.79	0.74	58
Happy	0.70	0.73	0.71	546
Sad	0.87	0.87	0.87	570
Surprise	0.87	0.89	0.88	189

Fig. 8. Classification Result Table

B. Conclusion

In this paper we proposed a modified Sequential Deep Learning Architecture with integrated probabilistic reasoning to address the challenges of Multilabel audio classification on the TESS, SAVEE, RAVDESS and CREMA-D datasets. Our approach shows notable improvement in classifying multiple emotions demonstrated by evaluation metrices such as accuracy, loss function, precision, recall and f1-score. The application of Dropping Layers after every convolution layer contributed to decrease in the overfitting of the model which helped in achieving such high accuracy and validation .The high accuracy achieved across multiple classes demonstrated that our approach can classify multiple audios in real-world applications.

REFERENCES

[1] H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9, doi: 10.1109/WACV.2016.7477679.

[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[3] Dias Issa, M. Fatih Demirci, Adnan Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, Volume 59, 2020, 101894, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2020.101894.

[4] J. F. Cohn et al., "Detecting depression from facial actions and vocal prosody," 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, Netherlands, 2009, pp. 1-7, doi: 10.1109/ACII.2009.5349358.

[5] Veena Potdar, Supritha M Bhatt, Yashaswini M K, Lavanya Santosh, "Analysis of Vocal Pattern to Determine Emotions using Machine Learning," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT), Volume 10, Issue 11 (November 2021).

[6] Andy Sun, Maisy Wieman, "Analyzing Vocal Patterns To Determine Emotion," http://www.datascienceassn.org/sites/default/files/Analyzing

[7] Wang, Chunyi, Ren, Ying, Zhang, Na, Cui, Fuwei, Luo, Shiying, "Speech emotion recognition based on multi-feature and multi-lingual fusion," Multimedia Tools and Applications, vol. 81, 2022, doi: 10.1007/s11042-021-10553-4.

[8] Abdullah, Sharmeen, Ameen, Siddeeq, M. Sadeeq, Mohammed, Zeebaree, Subhi, "Multimodal Emotion Recognition using Deep Learning," Journal of Applied Science and Technology Trends, vol. 2, pp. 52-58, 2021, doi: 10.38094/jastt20291.

[9] Steven R. Livingstone, 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, 2018, e0196391.

[10] Li Y, Tao J, Chao L, et al., "CHEAVD: A Chinese natural emotional audio–visual database," Journal of Ambient Intelligence and Humanized Computing, 2016, 8(6).

[11] Y. Kim, H. Lee and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 3687-3691, doi: 10.1109/ICASSP.2013.6638346.

[12] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII), Sep. 2015, pp. 827-831.

[13] Seokho Kang, "On Effectiveness of Transfer Learning Approach for Neural Network-Based Virtual Metrology Modeling," IEEE Transactions on Semiconductor Manufacturing, vol. 31, pp. 1-1, 2017, doi: 10.1109/TSM.2017.2787550.

[14] Bilal, M., Mughal, A., Fatima, T., Syed, Z.S., Syed, A. Mahboob, H., "Speech Emotion Recognition in Pakistani-Accented English Language," 2023 17th International Conference on Open Source Systems and Technologies, ICOSST 2023 - Proceedings, IEEE, Lahore, Pakistan, 2023, doi: 10.1109/ICOSST60641.2023.10414231.

[15] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, Guan-Zheng Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," Neurocomputing, Volume 273, 2018, Pages 271-280, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2017.07.050.

[16] Tharian, J., Nandakrishnan, R., Sajesh, S., Arun, A.V., and Jayadas, C.K., "Automatic Emotion Recognition System using tinyML," 2022 International Conference on Futuristic Technologies (INCOFT), IEEE, 2022, pp. 1-4.

[17] Keshun, Y., Puzhou, W., Peng, H., and Yingkui, G., "A sound-vibration physical-information fusion constraint-guided deep learning method for rolling bearing fault diagnosis," Reliability Engineering System Safety, 2024, p.110556.

[18] Cardoso, Pedro J.S., Rodrigues, J. M. F., "A framework for emotion and sentiment predicting supported in ensembles," Universidade do Algarve, Instituto Superior de Engenharia, 2022.

[19] G. Trigeorgis, F. Ringual, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, 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), IEEE, 2016, pp. 5200-5204.

[20] Ephrem Afele Retta, Eiad Almekhlafi, Richard Sutcliffe, Mustafa Mhamed, Haider Ali, and Jun Feng, "A New Amharic Speech Emotion Dataset and Classification Benchmark," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 1, Article 20, 2023, https://doi.org/10.1145/3529759.

[21] S. S. Stevens, J. Volkmann, E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," Journal of the Acoustical Society of America, vol. 8, no. 3, 1937, pp. 185-190, https://doi.org/10.1121/1.1915893.

[22] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, L.-H. Cai, "Music type classification by spectral contrast feature," 2002 IEEE International Conference on Multimedia and Expo, ICME'02 Proceedings, vol. 1, IEEE, 2002, pp. 113-116.

[23] Davis, S. B., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, 1980, pp. 65-74.