Improving Speech Emotion Recognition: A Semi-Supervised Approach for Fine-Grained Analysis

Kshitij Goyal*, Ishwak Sharda*, Amir Shabani
School of Computing
University of the Fraser Valley, BC
{kshitij.goyal, ishwak.sharda}@student.ufv.ca,
{amir.shabani}@ufv.ca

Abstract

A key challenge in speech emotion recognition (SER) is the lack of fine-grained datasets containing both emotion and intensity labels, which limits the performance of data-demanding deep learning models in applications like social companion robots. Most existing datasets cover only basic emotions and rarely include nuanced intensity annotations. To address this gap, we present using semi-supervised learning (SSL) to create a larger fine-grained SER (FGSER) dataset from limited available datasets. Our model classifies 5 distinct emotions—anger, sadness, happiness, disgust, and fear—each represented across three intensity levels: low, medium, and high. We propose two SSL approaches tailored to different application needs: a Random Forest Classifier (RFC) for edge-computing environments that demand computational efficiency, and a Convolutional Neural Network (CNN) for scenarios where higher accuracy is critical. Including only high-confidence predictions to the original small dataset will increase the size of the dataset and hence improvement of the classifier's accuracy and generalization. This enhancement supports the development of conversational AI with high emotional intelligence (EQ), advancing FGSER for richer human-computer and human-robot interactions, more specifically for social companion robotic applications.

1 Introduction

Emotions are integral to human communication, conveying information about an individual's mental state, intentions, and personality [1]. Speech Emotion Recognition (SER) aims to identify and classify emotions from speech signals, independent of semantic content [2]. SER plays a vital role in human-computer interaction (HCI), enabling systems to respond empathetically, with applications in security, healthcare, education, and customer interactions[2, 3].

Recognizing emotions accurately from speech remains challenging due to variability in expression, which includes differences in tones, pitches, and intensities [4]. Capturing fine-grained emotional intensities is essential, especially in applications like social robotics, where nuanced emotional responses are crucial for meaningful interaction. Despite these needs, existing SER datasets often lack sufficient annotations for emotion intensity, limiting their applicability for (fine-grained speech emotion recognition) FGSER tasks [5].

To address these limitations, we implement a semi-supervised learning (SSL) framework that leverages a small labeled dataset to annotate a larger set of unlabeled data. We develop two tailored approaches: one optimized for edge-computing scenarios with limited computational resources, and another designed for applications prioritizing high accuracy, where computational capacity is not a constraint. For edge applications, we employ an ensemble method using a Random Forest Classifier (RFC). In contrast, the high-accuracy model uses a Convolutional Neural Network (CNN).

The rest of this paper is organized as follows: In Section 2, we highlight the relevant research in this area followed by our methodology and experimental setup in Section 3. We then present and discuss the results in Section 4. Finally, Section 5, summarizes our contribution and highlights future works.

2 Related Work

Recent advancements in FGSER have employed various methodologies to capture nuanced emotional expressions. For instance, Wang et al. [6] introduced Speech

Emotion Diarization (SED), which identifies emotion classes and their temporal boundaries within utterances, thereby enhancing classification accuracy and boundary detection. However, SED lacks focus on intensity variations critical for FGSER, as noted by Hamza et al. [7]. Alternative approaches have utilized statistical models like Hidden Markov Models (HMMs) in speech emotion recognition, offering a probabilistic framework to model temporal dynamics in speech signals. Nonetheless, Song et al. [8] highlight that HMMs often struggle with capturing the complexity of fine-grained emotions due to their reliance on predefined states and transitions .

In recent years, deep learning techniques, particularly CNNs, have been effectively applied in speech emotion recognition tasks. For instance, Tang et al. [9] proposed a CNN-Transformer model with a multidimensional attention mechanism, achieving significant performance improvements. Similarly, Peng et al. [10] introduced an efficient speech emotion recognition model using multi-scale CNN and attention mechanisms, demonstrating enhanced accuracy. In addition to CNNs, LSTM networks have also been explored for speech emotion recognition due to their ability to capture temporal dependencies. Jafri et al. [11] demonstrated that combining CNNs with LSTMs improves recognition accuracy by leveraging both spatial and temporal features in speech data.

Furthermore, multimodal approaches have been explored to capture nuanced emotional cues. Li et al. [12] propose a multimodal approach for fine-grained speech emotion recognition, using temporal alignment mean-max pooling and a cross-modality excitement module to capture nuanced emotional cues across modalities. Their model outperforms baselines on real-world datasets, effectively enhancing prediction accuracy in fine-grained emotional recognition.

RFC classifiers have emerged as a viable alternative to CNNs for speech emotion recognition, particularly when computational resources and data availability are limited. Rezapour Mashhadi and Osei-Bonsu [2] demonstrated that RFCs are faster and more suitable for smaller datasets. Additionally, their computational efficiency and adaptability make them especially advantageous for real-world applications, as noted by Aishwarya et al. [13] in their exploration of efficient machine learning classifiers and ensemble methods for emotion recognition tasks . Despite these advancements, limited labeled data remains a primary obstacle in FGSER. Zhu and Sato [14] show that ensemble-based semi-supervised learning approaches like NST have the potential to expand labeled datasets with minimal manual annotation, as evidenced by experiments on the CREMA-D dataset.

Building upon these approaches, this work proposes a framework centered on RFC and CNN models within a semi-supervised learning paradigm to enhance FGSER accuracy and adaptability across intensity levels. This contributes to applications in HCI and social robotics, where nuanced emotional understanding is essential.

3 Methodology and Experimental Setup

In this section, we discuss the datasets, steps in data preparation, feature extraction, and the architecture and frameworks for SSL classification.

3.1 Datasets

There are several publicly available SER datasets with CREMA-D and RAVDESS being the ones with fine-grained labels but different levels. The CREMA-D dataset comprises 7,442 audio clips from 91 actors, with six emotions and four intensity levels (low, medium, high, unspecified)[15], providing fine-grained labels for our study. The RAVDESS dataset, containing 7,356 files from 24 actors with seven emotions, uses two intensity levels: normal and strong [16]. For consistency, we mapped ”normal” to ”medium” and ”strong” to ”high.”

3.2 Data Harmonization

To harmonize the datasets, we standardized the naming conventions and resampled all audio files to 16 kHz. Padding was applied to a maximum length of 220 frames, corresponding to the longest audio sample in our dataset, to ensure uniform input size for feature extraction and model training. This approach aligns with practices recommended in speech processing literature [17]. Furthermore, we focused on five emotions common across all datasets and categorized them into three intensity levels: low, medium, and high.

3.3 Data Augmentation

To address the limited dataset size and enhance model robustness, we applied data augmentation by adding Additive White Gaussian Noise (AWGN) at a signal-to-noise ratio (SNR) of 20 dB to each audio file. This approach is commonly used in speech processing to simulate real-world noise conditions and improve generalization [17]. Additionally, both models incorporate pitch-shift augmentation, which modifies the pitch of audio signals without affecting the tempo [18].

These techniques are commonly used in speech processing to improve generalization and robustness of models.

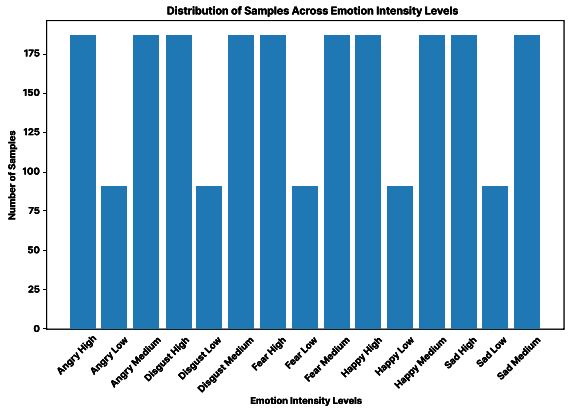

Figure 1: Distribution of emotion intensity levels after merging CREMA-D and RAVDESS datasets. Mapping RAVDESS intensity 0 to "medium" and 1 to "high" resulted in roughly 50% fewer "Emotion Low" samples compared to "Medium" and "High", resulting in an unbalanced dataset.

3.4 Data Splitting

We utilized the CREMA-D dataset, comprising 7,442 samples, and the RAVDESS dataset, which contains 1,440 samples. Since our focus is on fine-grained emotion classification, we excluded samples labeled as neutral, calm, and surprise, as they either lack intensity levels or do not align with our classification goals. After this filtering, only 960 samples from RAVDESS were retained. Additionally, some samples in the CREMA-D dataset lack defined intensity levels, so we filtered the dataset to include only those with specified intensities, resulting in 1,365 samples. This process yielded a combined dataset of 2,325 samples.

CREMA-D includes three intensity levels: high, mid, and low, while RAVDESS contains only high and mid intensity levels. To address this class unbalance, we applied stratified sampling [19] to ensure a balanced class distribution across both datasets. The resulting 2325 were further augmented using the augmentation techniques defined above, enhancing the model's robustness and diversity in data representation.

The dataset is then split into 80% for training and 20% for testing.

3.5 Feature Extraction

In alignment with existing literature, we use mel-spectrograms as our primary feature representation due to their effective capture of both spectral and temporal aspects essential for identifying different emotions and intensities in speech [20]. Mel-spectrograms align closely with human auditory perception, as they map frequencies to the mel scale, thereby providing a perceptually meaningful representation where lower frequencies are in narrower intervals, and higher frequencies are spaced wider apart [20]. Recent studies have shown mel-spectrograms to be highly effective in SER tasks [21]. Our preliminary experiments confirmed that mel-spectrograms achieved higher classification accuracy compared to other relevant features; Mel Frequency Cepstral Coefficients (MFCCs) and spectral contrast. Hence, for conciseness, we only include the mel-spectrogram in this paper. Using the librosa library [22], we computed mel-spectrograms with 128 mel bands, a frame size of 2048 samples, and a hop length of 512 samples. We converted power spectrograms to a decibel scale with *librosa.power-to-db* to compress the dynamic range and highlight perceptually relevant features.

3.6 Model Architecture and Training

3.6.1 Random Forest Classifier

As an ensemble-based classifier, RFCs are well-suited for edge computing environments due to their inherent parallelism and computational efficiency. Each tree in a random forest operates independently, allowing for parallel processing and efficient utilization of limited computational resources. This characteristic makes RFCs advantageous for deployment on edge devices, where computational power and memory are often constrained. Additionally, RFCs are robust against overfitting, as they aggregate multiple decision trees trained on different data subsets, enhancing generalization performance. This robustness is particularly beneficial when working with smaller or moderately sized datasets, such as those commonly found in FGSER tasks. The strengths of Random Forests, including their efficiency and effectiveness in various applications, have been well-documented [23].

The RFC model in this study is configured with 500 decision trees (estimators), each trained independently on bootstrapped subsets of the data to enhance generalization. To address class imbalance, the class_weight='balanced' parameter dynamically adjusts weights based on class frequencies in the input data. Default parameter settings in Python 3.9.19 were utilized for the number of features considered (max_features) and tree depth (max_depth)

3.6.2 Convolutional Neural Network

As a deep learning model, CNNs are highly effective for tasks requiring detailed feature extraction, such as SER. They excel at capturing intricate temporal and spectral patterns in audio spectrograms, which are essential for distinguishing nuanced emotions and their intensities. CNNs can even directly model raw speech signals, ef-

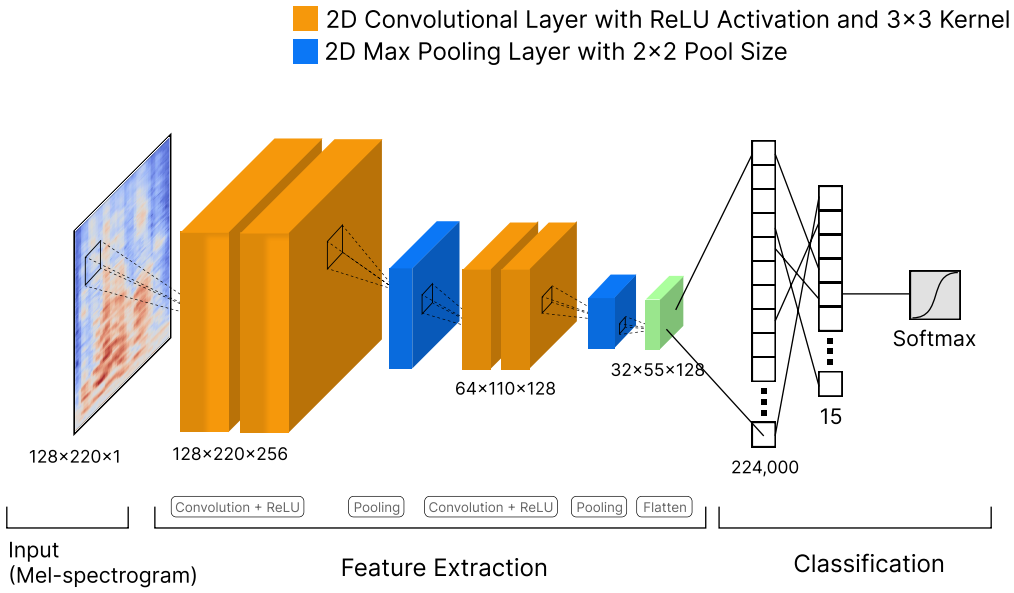

Figure 2: Architecture of 2D-CNN

fectively learning complex representations for emotion recognition [24].

However, this enhanced accuracy comes with increased dataset size and computational demands, making CNNs more suitable for centralized systems with substantial resources rather than edge devices with limited capabilities and small dataset. Studies on efficient CNN architectures have explored methods to reduce computational load while maintaining performance, highlighting the trade-offs between accuracy and resource consumption in deploying CNNs for SER[25].

Our CNN model, shown in Figure 2, consists of four convolutional layers with progressively smaller filter sizes, optimized for high-dimensional datasets like CREMA-D. Each convolutional layer uses a ReLU activation function and is followed by MaxPooling to reduce dimensionality. Dropout layers, with values of 0.1 and 0.5, are included at appropriate stages to prevent overfitting and improve generalization. The model also integrates L1 and L2 regularization across its convolutional and dense layers to further control overfitting. The network concludes with a fully connected dense layer with 15 units (representing emotion classes) and a softmax activation for classification. The Adam optimizer[26] is used for training, with a learning rate decay of 10^{-6}, ensuring smooth convergence.

3.7 Semi-Supervised Learning (SSL)

Our initial combined RAVDESS-CREMA-D dataset includes 2,325 samples, of which 80% (1,860 samples) are used for training. Neutral samples were removed from the unlabeled list in CREMA-D, resulting in 4,990 unlabeled samples for prediction. These samples are annotated with emotion type (e.g., anger) but lack intensity-level labels (e.g., high/low). Following prediction, we retained only high-confidence samples, discarding those with incorrect emotion type predictions regardless of intensity level. In the SSL framework (see Fig. 3), these samples are then assigned pseudo-labels and added to create a larger dataset, without manual annotation [27], for re-training to enhance the model performance.

4 Results

This section presents the results of two experimental stages using both RFC and CNN models. The first experiment uses an initial dataset containing 1,860 training samples (referred to as Training Set 1 in the tables). In the second experiment, we expand this set by incorporating additional pseudo-labeled samples generated via SSL to create Training Set 2. For testing, we use 4,990 unlabeled samples from the CREMA-D dataset. These samples, which lack intensity labels but have emotion type labels, are passed to RFC or CNN models for emotion type and intensity predictions. Only samples with correct emotion type predictions are accepted and assigned a pseudo-label. To evaluate the impact of increased dataset size on model performance, we then add these pseudo-labeled samples to the original dataset for re-training and testing creating Training Set 2. We

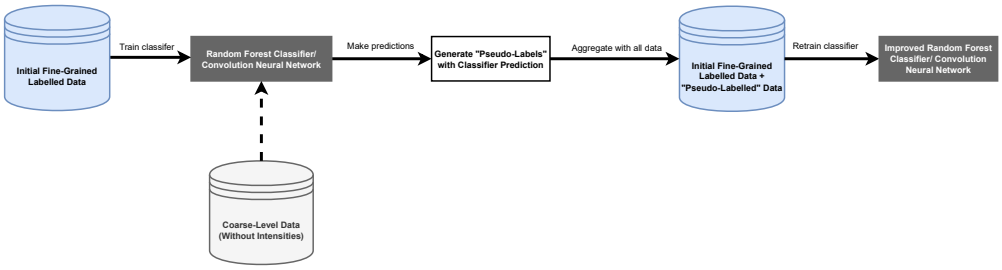

Figure 3: SSL Framework for Creating a Larger FGSER Dataset.

chose four metrics—accuracy, precision, recall, and F1-score—because they provide complementary insights. Accuracy measures overall correctness, precision evaluates the model's reliability for specific classes, recall ensures sensitivity to all instances, and the F1-score balances the trade-off between precision and recall.

Out of the 4,990 unlabeled samples, the CNN model accurately predicted the emotion type for 1,703 samples (around 30%), allowing us to accept the corresponding intensity level predictions and expand the dataset for retraining. Following this increase, CNN performance improved, with average accuracy rising from 82% to 85%, demonstrating the model's enhanced performance with a larger dataset—a common advantage in deep learning.

On the other hand, RFC accurately predicted the emotion type of approximately 1990 samples (around 40%). As an ensemble model, RFC's performance remained statistically consistent, demonstrating its robustness with smaller datasets and its suitability for edge computing applications. Notably, the CNN achieved about 16% higher accuracy than RFC, suggesting that extending our SSL approach to include additional publicly available datasets could further enhance the accuracy of our deep learning model. Moreover, an analysis of the confusion matrices in Figure 4 and Figure 5 reveals the models' performances on specific emotion-intensity pairs. For instance, the RFC model struggles to differentiate Fear_Low from Sad_Medium due to overlapping spectral features. In contrast, the CNN achieves better differentiation across intensity levels, as reflected in reduced misclassifications in its confusion matrix.

Table 1: The average performance of RFC model is statistically the same with the increase training dataset size. Training Set 1 contains 1,860 samples and Training Set 2 contains 3,452 samples.

Metric	With Training Set 1	With Training Set 2
Accuracy	69.11% ± 1.03%	**69.97% ± 1.68%**
Precision	68.47% ± 0.83%	69.73% ± 1.94%
Recall	68.93% ± 1.16%	67.67% ± 1.54%
F1-Score	68.00% ± 1.13%	68.00% ± 1.65%

Table 2: The average performance of CNN model increases with the training dataset size. Training Set 1 contains 1,860 samples and Training Set 2 contains 3222 samples.

Metric	With Training Set 1	With Training Set 2
Accuracy	82.00% ± 1.46%	**85.00% ± 1.15%**
Precision	81.50% ± 1.41%	84.31% ± 1.25%
Recall	81.31% ± 1.62%	84.25% ± 1.18%
F1-Score	81.13% ± 1.54%	84.32% ± 1.20%

5 Conclusion

In this paper, we proposed a SSL approach to expand FGSER datasets by predicting emotion intensity levels for unlabeled samples, thereby enhancing model accuracy. Two models were evaluated: a RFC classifier and a CNN. The RFC, chosen for its computational efficiency and suitability for edge computing, demonstrated stable performance with limited data. The CNN, by contrast, showed improved accuracy with the additional pseudo-labeled samples, underscoring the data scalability advantages of deep learning models. Overall, our SSL method successfully increased dataset size, contributing to improved performance of deep learning models, especially in data-intensive FGSER tasks. These advancements are particularly valuable in social companion robotics, where a nuanced understanding of emo-

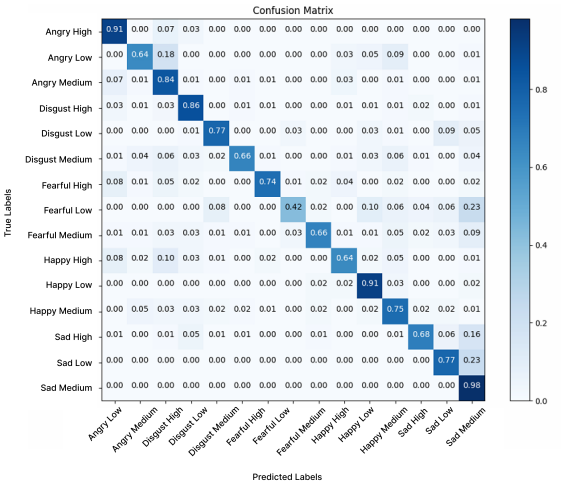

Figure 4: Confusion Matrix with Training Set 2 for RFC. The main confusion is between Fear Low and Sad Medium due to overlapping spectral features.

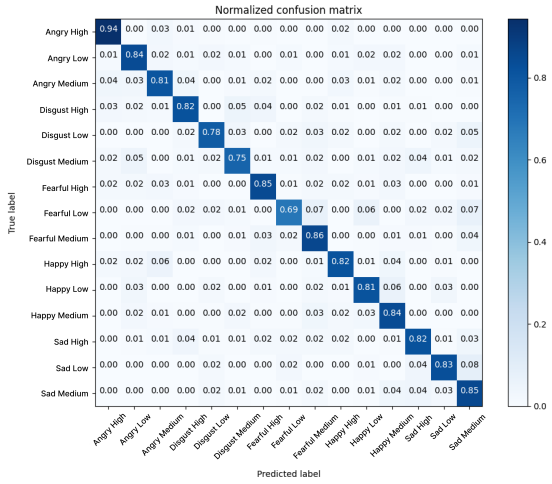

Figure 5: Confusion Matrix with Training Set 2 for CNN shows significantly less confusion with other classes.

tions is crucial for fostering meaningful HCI.

Looking ahead, we plan to extend our SSL approach to additional datasets, such as SAVEE[28] and TESS[29], to further enhance dataset size, diversity, and model robustness. We also aim to conduct comparative studies leveraging transfer learning with deep learning architectures like ResNet, AlexNet, and InceptionNet, which have demonstrated strong performance in emotion recognition tasks[30]. Additionally, we intend to explore models specifically tailored for signal processing and emotion classification, such as LSTMs[11] and MAMBA [31], which excel at capturing temporal and contextual dependencies in speech data. By combining the advantages of transfer learning and these state-of-the-art architectures, we anticipate further improvements in both model performance and generalization.

6 Acknowledgments

The authors gratefully acknowledge the financial support provided by the Esposito Family Center for Innovation and Entrepreneurship (EFCIE) and TD Bank's Better Health - Innovative Solutions program for this research project.

References

[1] J. Dennison, "Emotions: functions and significance for attitudes, behaviour, and communication," *Migration Studies*, vol. 12, no. 1, pp. 1–20, 08 2023.

[2] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," *PLOS ONE*, vol. 18, no. 11, pp. 1–13, 11 2023. [Online]. Available: https://doi.org/10.1371/journal.pone.0291500

[3] S. Ahuja and A. Shabani, "Affective computing for social companion robots using fine-grained speech emotion recognition," pp. 331–332, 2023.

[4] A. Schirmer and R. Adolphs, "Emotion perception from face, voice, and touch: Comparisons and convergence," *Trends in Cognitive Sciences*, vol. 21, no. 3, pp. 216–228, Mar 2017.

[5] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data augmentation and deep learning methods in sound classification: A systematic review," *Electronics*, vol. 11, no. 22, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/22/3795

[6] Y. Wang, M. Ravanelli, and A. Yacoubi, "Speech emotion diarization: Which emotion appears when?" 2023. [Online]. Available: https://arxiv.org/abs/2306.12991

[7] H. Hamza, F. Gafoor, F. Sithara, G. Anil, and V. S. Anoop, "Emodiarize: Speaker diarization and emotion identification from speech signals using convolutional neural networks," 2023. [Online]. Available: https://arxiv.org/abs/2310.12851

[8] M. Song, C. Chen, J. Bu, and M. You, "Speech emotion recognition and intensity estimation," in *Computational Science and Its Applications – ICCSA 2004*, A. Laganá, M. L. Gavrilova, V. Kumar, Y. Mun,

C. J. K. Tan, and O. Gervasi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 406–413.

[9] X. Tang, Y. Lin, T. Dang, Y. Zhang, and J. Cheng, "Speech emotion recognition via cnn-transformer and multidimensional attention mechanism," 2024. [Online]. Available: https://arxiv.org/abs/2403.04743

[10] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2021, p. 3020–3024. [Online]. Available: http://dx.doi.org/10.1109/ICASSP39728.2021.9414286

[11] B. Kaushik, "A hybrid technique using cnn+lstm for speech emotion recognition," *International Journal of Engineering and Advanced Technology*, vol. 9, pp. 1126–1130, 08 2020.

[12] H. Li, W. Ding, Z. Wu, and Z. Liu, "Learning fine-grained cross modality excitement for speech emotion recognition," 2021. [Online]. Available: https://arxiv.org/abs/2010.12733

[13] N. Aishwarya, K. Kaur, and K. Seemakurthy, "A computationally efficient speech emotion recognition system employing machine learning classifiers and ensemble learning," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 239–254, 2024. [Online]. Available: https://doi.org/10.1007/s10772-024-10095-8

[14] Z. Zhu and Y. Sato, "Speech emotion recognition using semi-supervised learning with efficient labeling strategies," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 358–365.

[15] 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, Oct-Dec 2014.

[16] 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, pp. 1–35, 05 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391

[17] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," pp. 1089–1093, 2017.

[18] S. Wei, S. Zou, F. Liao, and weimin lang, "A comparison on data augmentation methods based on deep learning for audio classification," *Journal of Physics: Conference Series*, vol. 1453, no. 1, p. 012085, jan 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1453/1/012085

[19] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset," *Electronics*, vol. 12, no. 21, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/21/4423

[20] H. Li, J. Li, H. Liu, T. Liu, Q. Chen, and X. You, "Meltrans: Mel-spectrogram relationship-learning for speech emotion recognition via transformers," *Sensors*, vol. 24, no. 17, 2024.

[21] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," pp. 1–6, 2019.

[22] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and Music Signal Analysis in Python," pp. 18 – 24, 2015.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[24] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *CoRR*, vol. abs/1904.03833, 2019. [Online]. Available: http://arxiv.org/abs/1904.03833

[25] C. Huang, "Ringcnn: Exploiting algebraically-sparse ring tensors for energy-efficient cnn-based computational imaging," *CoRR*, vol. abs/2104.09056, 2021. [Online]. Available: https://arxiv.org/abs/2104.09056

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[28] P. Jackson and S. ul haq, "Surrey audio-visual expressed emotion (savee) database," 04 2011.

[29] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[30] M. Jakubec, E. Lieskovska, R. Jarina, M. Spisiak, and P. Kasak, "Speech emotion recognition using transfer learning: Integration of advanced speaker embeddings and image recognition models," *Applied Sciences*, vol. 14, no. 21, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/21/9981

[31] X. Li, X. Fan, Q. Wu, X. Peng, and Y. Li, "Mamba-enhanced text-audio-video alignment network for emotion recognition in conversations," 2024. [Online]. Available: https://arxiv.org/abs/2409.05243