Literature Review

[1] Liu et al.(2023)[1] proposed that Speech Emotion Recognition (SER) encounters significant challenges due to limited datasets and the subjective nature of emotion perception, with many existing models relying heavily on techniques from computer vision and natural language processing. To address these issues, we propose a novel approach that draws inspiration from human emotional perception, incorporating implicit emotional attribute classification through multi-task learning. This method aims to enhance the recognition and classification of emotions in speech. Our experiments utilize the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, leveraging deep learning algorithms within a multi-task learning framework to improve both unweighted and weighted accuracy in SER tasks

[2] Speech Emotion Recognition (SER) struggles with high classification accuracy due to challenges like speaker variability, class imbalance, and the nuanced nature of emotional expression, particularly in Speaker-Independent scenarios are proposed in Madanian et al.,2023[2].Also, this paper offers a comprehensive framework that enhances SER through advanced feature extraction and classification techniques, supported by robust data preprocessing and diverse datasets. We utilize Emo-DB, RAVDESS, and CREMA-D, which provide a variety of emotional expressions across different speakers, improving model training. The study employs Support Vector Machines (SVM) and Random Forests for traditional classification, alongside deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to capture complex features and temporal patterns in speech.

[3] This research addresses the limitations of traditional Mel Frequency Cepstral Coefficients (MFCCs) in Speech Emotion Recognition (SER), which often struggle to capture the necessary temporal dynamics for accurate emotion classification. To enhance performance, a hybrid feature extraction approach is proposed, combining MFCCs with time-domain features (termed MFCCT). This combined feature set is then utilized within a Convolutional Neural Network (CNN) framework. The study employs three widely recognized datasets: Emo-DB, SAVEE, and RAVDESS, each providing diverse emotional speech data. The MFCCT features, in conjunction with the CNN, significantly improve classification accuracy over conventional methods, demonstrating the effectiveness of hybrid approaches in SER tasks were the views shared in Alluhaidan et al.,2023

[4] With an emphasis on problems including feature variability, classification accuracy, and scarce datasets, this work addresses important issues in speech emotion recognition (SER) systems. It presents a revolutionary method for producing a more accurate depiction of emotions by combining linguistic, prosodic, and spectral elements. The study combines deep learning techniques like Convolutional Neural Networks and Recurrent Neural Networks with conventional machine learning methods like Support Vector Machines and Random Forests using popular datasets like EMO-DB, RAVDESS, and TESS. This comprehensive approach seeks to improve emotion classification's resilience and accuracy while advancing SER systems.Survey on speech emotion recognition : Features,classification schemes and databases EL ayadi M ,kamel m.s,karray F. pattern recognition(2011)

[5] This paper addresses the challenge of Speech Emotion Recognition (SER), where emotional expression in speech varies significantly across individuals, contexts, and environments, making accurate detection difficult. To tackle this, we propose a hybrid CNN-LSTM model that extracts and classifies emotional features from speech, leveraging MFCC and pitch as key features. The model is trained on publicly available datasets like Emo-DB, RAVDESS, and TESS, which provide diverse emotional speech samples. Principal Component Analysis (PCA) is used for feature selection, reducing dimensionality and improving model efficiency. This hybrid approach, combining CNN's spatial feature extraction and LSTM's temporal processing, aims to enhance the model's ability to generalize across varying emotional expressions and real-world noise, providing a robust solution for real-time emotion detection from speech(SPEECH EMOTION RECOGNITION USING CNN-LSTM Ms Gayathri R 1, Arun Kumar B2, Inbanathan S3,Karthick S4 **)**

[6] Speech Emotion Recognition (SER) is challenging due to the complex and variable nature of emotional expression in speech. To address this, we propose a novel approach that utilizes Deep Neural Networks (DNNs) for automatic feature extraction from raw speech, generating segment-level emotion probability distributions. These segment-level features are then aggregated into utterance-level features and classified using an Extreme Learning Machine (ELM), a fast and efficient algorithm. This combination allows the model to capture complex emotional patterns and provide accurate classification with a significant performance boost of 20% relative accuracy compared to existing methods. The approach is validated using publicly available datasets like EmoDB, RAVDESS, and TESS, demonstrating its effectiveness in recognizing emotions from speech.(Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine Kun Han1, Dong Yu2, Ivan Tashev2)

[7] Emotion Speech Recognition (ESR) systems often focus on feature extraction, neglecting the crucial classification stage, leading to suboptimal performance. This research addresses this gap by proposing the use of an Optimized Genetic Algorithm-Extreme Learning Machine (OGA-ELM) for classification, combined with Mel Frequency Cepstral Coefficients (MFCC) for feature extraction. The system is evaluated using the Berlin Emotional Speech (BES) dataset, which includes seven emotions, across four scenarios: Subject Independent (SI), Subject Dependent (SD), Gender Dependent Female (GD-Female), and Gender Dependent Male (GD-Male). The OGA-ELM optimizes ELM through a Genetic Algorithm (GA), enhancing classification accuracy and reducing training time. This approach demonstrates improved performance and fast execution, making it well-suited for real-time emotion recognition applications.(Speech emotion recognition using optimized genetic algorithm-extreme learning machine [Albadr M. A. A.](https://www.mendeley.com/authors/57201726613)[Tiun S.](https://www.mendeley.com/authors/24825640000)[Maen M. K.](https://www.mendeley.com/authors/57491490900))

[8] Speech Emotion Recognition (SER) focuses on detecting emotions from speech but encounters challenges like variations in speaker expressions, the need for real-time processing, and limited labeled datasets. To overcome these hurdles, a combination of deep learning models (such as CNNs and RNNs) and traditional machine learning methods is suggested, along with data augmentation and transfer learning techniques to boost performance. Datasets like EMO-DB, RAVDESS, TESS, and CREMA-D are utilized for training models with diverse emotional content. Algorithms including Support Vector Machines (SVM), Random Forests, and LSTMs are commonly applied to enhance the accuracy and effectiveness of emotion recognition, facilitating improved interactions between humans and machines.(Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models)

[9] Speech Emotion Recognition (SER) seeks to identify human emotions from verbal expressions but faces challenges such as variability in speech, the need for real-time processing, and limited diverse datasets. Proposed solutions include hybrid approaches that combine deep learning models (like CNNs and LSTMs) with traditional machine learning techniques (such as SVMs and Random Forests), as well as data augmentation and transfer learning to enhance model performance. Commonly used datasets include EMO-DB, RAVDESS, TESS, and CREMA-D, which provide varied emotional content. Algorithms employed in SER span deep learning and traditional machine learning, enabling more effective emotion classification and improving human-computer interaction.[**Speech emotion recognition using neural networks**](https://www.mendeley.com/catalogue/bffde654-bb45-3475-b5c7-32e988a97e7d)

[10] Speech Emotion Recognition (SER) seeks to classify human emotions from verbal expressions but faces challenges such as speech variability, the need for real-time processing, and limited diverse datasets. This study proposes a new architecture that extracts features including mel-frequency cepstral coefficients, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast directly from raw audio data, utilizing a one-dimensional Convolutional Neural Network (CNN) for emotion identification. The research employs the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Berlin Emotional Database (EMO-DB), and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. By implementing an incremental approach to improve classification accuracy, the model effectively classifies emotions without converting audio to visual representations, enhancing the robustness of SER systems.([Speech emotion recognition with deep convolutional neural networks](https://www.mendeley.com/catalogue/fcf3a9c6-3e38-3b2c-a745-161269dc8643))

[11] Speech Emotion Recognition (SER) seeks to identify and classify human emotions from speech signals but faces challenges such as speaker variability, background noise, and the complexity of emotional expression. This study proposes a framework that extracts both spectral and prosodic features, including Mel-frequency cepstral coefficients (MFCCs), fundamental frequency, loudness, pitch, and glottal parameters, to enhance emotion recognition. The research utilizes datasets such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin Emotional Database (EMO-DB), and Interactive Emotional Dyadic Motion Capture (IEMOCAP). Key algorithms employed include Support Vector Machines (SVM) for gender classification and Radial Basis Function (RBF) Networks for emotion recognition, with RBF networks demonstrating superior accuracy compared to Back Propagation Networks.([Human speech emotion recognition](https://www.mendeley.com/catalogue/74286974-70a0-3904-b202-3bd877f67cc9))

[12]  This study aims to develop an effective machine learning model that analyzes speech features, which are commonly impaired in PD patients, to predict the severity of the disease. The research utilizes the Parkinson’s Telemonitoring dataset from the UCIML repository, containing speech samples and relevant features, to train and evaluate the model. Various machine learning algorithms, including Support Vector Machines (SVM) and Radial Basis Function (RBF) Networks, are employed for classification tasks, chosen for their effectiveness in analyzing speech data and accurately predicting PD severity based on extracted features.([Prediction of parkinson’s disease using machine learning techniques on speech dataset](https://www.mendeley.com/catalogue/ade4bb55-09be-3dc6-bd84-9284a191cd1f))

[13] The rise of social media has significantly increased the prevalence of hate speech, posing substantial challenges for online platforms. Current detection methods often rely on subjective judgments, which can lead to inefficiencies and inconsistencies. This highlights the urgent need for automated systems that can accurately identify and categorize hate speech messages in real time. This study seeks to evaluate the performance of three feature engineering techniques alongside eight machine learning algorithms using a publicly available dataset that encompasses three distinct categories of hate speech. By examining approaches such as Bag of Words, TF-IDF, and n-grams in combination with algorithms like Support Vector Machines (SVM), Logistic Regression, Decision Trees, and Random Forests, the research aims to identify the most effective strategies for automated hate speech detection.([Automatic hate speech detection using machine learning: A comparative study](https://www.mendeley.com/catalogue/ed73345c-9e4b-3ff6-bda0-7712656062c9))

[14]This study addresses the challenge of accurately recognizing emotions from speech, which is essential for enhancing human-computer interaction. To tackle this issue, the research proposes a method that utilizes Support Vector Machines (SVM) with Mel Frequency Cepstral Coefficients (MFCC) extracted from voice signals to classify five specific emotions: anger, happiness, neutral, pleasant surprise, and sadness. The methodology is evaluated using two datasets: the Toronto Emotion Speech Set (TESS) and the Interactive Emotional Dyadic Motion Capture (IEMOCAP), both of which provide diverse emotional speech samples. The algorithms employed include SVM-Gaussian and SVM-Quadratic, yielding high accuracy rates of 97% for the TESS dataset and 86% for the IEMOCAP dataset, demonstrating the effectiveness of the proposed approach in emotion recognition tasks.([MFCC and Machine Learning Based Speech Emotion Recognition Over TESS and IEMOCAP Datasets](https://www.mendeley.com/catalogue/9e6c2ca7-bd12-3f6d-ab6d-6c9dcc5be5b5))