Review on Emotion-Based Speech Analysis For Disaster Response and Crisis Management

Rafa Siddiqua
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rafa.siddiqua@g.bracu.ac.bd

Rabea Akhter
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rabea.akhtar@g.bracu.ac.bd

Samiu Mostafa Ishan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
samiu.mostafa.ishan@g.bracu.ac.bd

Sania Azhmee Bhuiyan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sania.azhmee.bhuiyan@g.bracu.ac.bd

Farah Binta Haque
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
farah.binta.haque@g.bracu.ac.bd

Adib Muhammad Amit
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
adib.muhammad.amit@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—This review paper gives an exhaustive investigation and blend of ongoing progressions and systems utilized in emotion recognition by speech which is the way to human feeling communication. Nonetheless, the nonlinear qualities of emotion are variable, complex, and unobtrusively evolving. As a result, emotion recognition from speech remains difficult. This review paper has utilized various methodologies, and this research focuses on the qualities and limits of various examinations, featuring arising examples, irregularities, and regions requiring further research. The latest Discourse Emotional Acknowledgment SER (Speech emotion recognition), IMEMD-CRNN (Integrated Multimodal Emotion Detection with Convolutional Recurrent Neural Networks), and CNN-LSTM (Convolutional Neural Networks and Long Short-Term Memory) procedures for emergencies are examined in this theory. Approaches, for instance, planning BiLSTM (Bidirectional Long Short-Term Memory) and LSTM (Long Short-Term Memory), CNN-LSTM associations, and involving the IMEMD-CRNN structure display promising movements in feeling affirmation. This review tries to explain the meaning of various combinations for upgraded emotional recognition using speech and gives a guide to the advancement of further developed emergency board models and designs from now on.**

***Index Terms*—Emotion recognition, Disaster, Crisis management, IMEMD-CRNN, CNN-LSTM, BiLSTM, CNN, LSTM.**

I. INTRODUCTION

In an environment of rapid technological advancement, understanding and analyzing human emotions in emergencies has become an important pursuit in many fields. This review is designed to provide comprehensive research and report on various aspects of emotional intelligence in stress, including emotional intelligence (SER), social-emotional analysis of energy during natural disasters [5], and the concept of signal communication in crises. detection [4] and sentiment analysis on platforms such as Twitter during critical events [1]. Speech Emotion Recognition (SER) and emergencies. In human-computer interaction (HCI), speech recognition plays an important role in this breakthrough. Thoughts expressed in words are especially important in emergencies [7]. Its applications include call centers, healthcare, and digital marketing. However, improving the accuracy of natural language emotion recognition is still a challenge. Research is constantly working towards the creation of systems that can overcome language barriers, recognize different views of the speaker, and work well in noisy environments [2]. The role of social media in disaster response. The emergence of social media platforms has revolutionized the dissemination of messages and information. Expression of emotions, especially during natural disasters [8]. This chapter highlights the importance of public opinion surveys in assessing public opinion, influencing the way public opinion is influenced, and influencing emergency decision-making, providing a good insight into disaster management [6]. Voice Analysis Research in CrisisIt is important to understand the emotions sent by voice signals in crises. Traditional methods use multitasking to identify emotions in the mind to help solve communication problems when sending voice data across the network in critical situations [4]. Sentiment analysis and discovery on Twitter. Twitter's immediacy makes it useful during important events, providing a quick understanding of public opinion and sentiment[1]. This chapter explores the integration of event detection and sentiment analysis using Twitter profiles, using the Las Vegas shooting study as an example [1]. This comprehensive review integrates and presents a wide range of research theories in crises and suggests that competition among leaders and misunderstandings of the response process are important for solving future problems in crisis management.

II. Aim and Objective

A. *Objective*

This research has the target to compare different emotion recognition by using speech and to figure out which is the best one by providing a brief comparison between different implementations, methodologies, and datasets. Provide a better view for the readers to make them understand what they can use or what will be the best way to use it.

B. *Aim*

- What is the outcome of the comparison between the provided models and their implementation?
- What is the outcome of the comparison between the results and figure out what will be the best to use according to their accuracy?
- To provide a final result for them.

III. Literature Review

To find out the impact of emotions in a speech, we have gone through many research papers and journals that have extensively talked about emotion patterns in speeches. Authors Sung-Woo Byun and Seok-Pil Lee made a database containing emotional speech that would analyze emotions related to speeches in the Korean language[1]. Authors Gang Liu, Shifang Cai, and Ce Wang had taken the approach of multitask learning, where they trained all the implicit attribute classification and speech-emotion classifiers at the same time. They also carried out a binary classification experiment of implicit emotion attributes, where the results justify the reliability of their hypothesis[2]. Authors Bagus Tris Atamaja, Kiyoaki Shirai, and Masato Akagi have proposed a methodology, where firstly they have used the approach of Speech-based Emotion Recognition with LSTM networks. After that, they used word embeddings-based Emotion Recognition with Dense networks. Finally, they have combined the two methods to form a Dense network to predict the recognition of emotions related to speeches[3]. Authors of the paper[4] have used a machine learning technique, where emotions are classified into five different categories, namely anger, fear, happiness, sadness, and disgust. They have used audio signals to extract features required to train classifiers that would easily recognize the underlying emotion related to the speech. Authors Anuja Thakur and Sanjeev Dhull[5] have talked about different approaches for developing speech recognition that would be independent of language and speaker. They have also talked about different pre-processing techniques, feature extraction methods, and classifiers used for speech emotion recognition. Authors of [8] have discussed the psychological impact of emotion on speech, where they have explored the human speech production system to detect the emotionally significant regions of speech. Authors of the paper[9] have used the sliding window method and an Artificial Neural Network(ANN) model to extract features for Speech Emotion Detection(SED).

IV. Proposed Methodology

Emotion-based speech analysis for disaster response and crisis management typically involves a combination of algorithms and techniques from natural language processing (NLP), machine learning, and signal processing. Consolidating discourse information with different modalities like looks, physiological signs (e.g., pulse, skin conductance), or printed content for better feeling acknowledgment [21]. Different AI (ML) and profound learning (DL) models are utilized for close-to-home discourse acknowledgment, including Backing Vector Machines (SVM) [19]. Irregular Backwoods, Intermittent Brain Organizations (RNNs), Convolutional Brain Organizations (CNNs), and Long Transient Memory Organizations (LSTMs) [20]. Author Tris, Sirai, and Massto have used different features to recognize the emotions from a different speech where they have used different features like extraction from speech, acoustic feature extraction, and speech emotion recognition models where the whole speech and voice segments using bidirectional LSTM networks, with or without attention models [6]. Here the author has utilized discourse Feeling Acknowledgment where acoustic elements are separated from discourse fragments after quietness expulsion [22]. Highlights incorporate time and ghostly area highlights, MFCCs (Mel-recurrence cepstral coefficients), and chromas. Alongside that, different profound learning designs (LSTM, consideration models) are assessed for discourse-based feeling acknowledgment. The author additionally utilized the word Implanting feeling acknowledgment where they separated printed information from records tokenized and changed over into word embeddings alongside various profound learning structures (CNN, LSTM, LSTM with consideration) are investigated for text-based feeling acknowledgment. Sensory recognition generally involves extracting various acoustic, prosodic, and spectral features from speech signals [17]. These may include voice, energy, structure, MFCC, prosody (intonation, rhythm), and spectral features [18]. In Consolidating Discourse and Text Highlights creator has proposed a methodology that includes joining the acoustic elements from discourse with word embeddings from text information with various models, including CNN, LSTM, and mixes of organizations, which are assessed for joined feeling acknowledgment [6]. As per authors Sun, Li, and Mama, they have zeroed in on a methodology called IMEMD-CRNN (Further developed Concealing Exact Mode Decay - Convolutional Repetitive Brain Organization) for foreseeing feelings in discourse signals. The strategy comprises of three fundamental modules: IMEMD-based profound discourse signal deterioration, extraction of time-recurrence highlights from IMFs (Inherent Mode Capabilities), and discourse feeling acknowledgment in light of CRNN (Convolutional Repetitive Brain Organization) [7]. Here creators have utilized different IMEMD-based Close-to-home Discourse Signal Decay like EMD (Observational Mode Disintegration), Covering Signal-based EMD (MSEMD) and Further developed Concealing EMD (IMEMD). In EMD they have utilized motional discourse signal deterioration Non-fixed signals are separated by this into IMFs (Natural Mode Capabilities) and a buildup likewise various addresses mode mixing issues in EMD by using a sinusoidal veiling sign to disconnect different repeat parts alongside that creator additionally proposes a unique procedure to fabricate disguising signs that relieve mode mixing. It adds a disguising sign to the principal sign, breaks down it using EMD, and dispenses with the veiling sign to get high-repeat parts [7]. Extraction of Highlights In light of IMEMD Tone Elements: uses IMEMD decay to separate Hilbert range dispersion and shape highlights, among other ghostly elements. Mel-repeat cepstral coefficients (SMFCC) from the recreated signal procured through IMEMD, close by the first and second auxiliaries of SMFCC for discovering passing information. Convolutional Discontinuous Cerebrum Association (CRNN) which contains four 2D CNN blocks, followed by bidirectional GRUs and related layers, utilizing softmax inception at the outcome layer, they have likewise utilized the Adam enhancer and cross-entropy misfortune to prepare the organization over different ages at a foreordained learning rate and little bunch size [7]. The overall objective of the proposed method is to use a CRNN architecture and IMEMD-based signal decomposition, followed by feature extraction, to boost the robustness and accuracy of speech-emotion recognition systems. The IMEMD strategy means to relieve mode blending in EMD, while the CRNN model is utilized to gain and order feelings from the removed elements. This approach is intended to improve feeling forecast in discourse and announces consolidating signal handling procedures with profound learning techniques. According to Vydana, P. Vikash, T. Vamsi, K. P. Kumar, and A. K. Vuppala Identifying emotionally Important Areas where one needs to calculate is utilized to recognize sincerely critical sections inside an expression [8]. These fragments address the durationally short emotive motions made by the speaker. Along with that highlight extraction of ghastly vectors when the genuinely huge districts are recognized, phantom vectors are processed from the discourse information inside these distinguished portions. Using model turn of events where Gaussian Blend Displaying (GMM) which strategy is utilized to make models for feeling acknowledgment utilizing the unearthly vectors extricated from the genuinely critical portions [8]. According to Liu, Y. Mou, Y. Ma, C. Liu and Z. Dai the review proposes a methodology for perceiving feelings in discourse through a sliding window-based technique combined with counterfeit brain organization (ANN) displaying. Where one needs to use non-stop sliding windows of a particular length (Mδ) and slide distance (Δ) to perceive feelings window by window inside a discourse test [9]. Identifying emotional significance to critical locales inside every expression utilizing sliding windows, creating a succession addressing close-to-home vectors. Also, weight conveyance capability and lattice development put a weight circulation capability to portray the commitment of feeling from span level to window-level feelings. Develop a framework (G) to plan span-level close-to-home vectors to window-level profound vectors. This models close-to-home dispersions inside every window utilizing Gaussian likelihood thickness capabilities, approximating profound commitment inside every window. Also, fusion and extraction of features from where one can determine span level close to home vectors (e*) given the weight dispersion and straightforwardly perceives window-level profound vectors. Both e* and E are considered as highlights and melded. End with testing the EMO-DB Dataset: Assesses the proposed model utilizing the Berlin Feeling Data set (Emotional DB), choosing explicit feelings for examination. Conducts five-overlay cross-approval for preparing and testing sets, guaranteeing speaker autonomy [9]. According to Sharma, Dutta, and Pradhan, coordinate discourse information with physiological signs like pulse, and skin conductance, or look to make a multimodal dataset. Using combination strategies like late combination (consolidating highlights at a later stage) or early combination (incorporating highlights at the info level) to thoroughly catch profound prompts more [12]. Consolidating logical data from the discourse content, environmental factors, or progressing occasions to all the more likely figure out the close-to-home state. Using context-oriented embeddings or consideration systems to gauge the significance of various logical components in feeling acknowledgment [13]. Utilizing pre-prepared models on huge close-to-home discourse datasets and calibrating them on unambiguous fiasco-related profound discourse datasets [14]. Using move figuring out how to conquer information shortage in calamity explicit situations. Growing continuous feeling discovery frameworks that can dissect progressing discourse information to give quick criticism or help in emergency board procedures. Using lightweight models upgraded for speed and exactness. Stretching out feeling-based discourse examination to different dialects pervasive in misfortune-impacted areas. Involves cross-lingual models or multilingual methodologies for feeling acknowledgment in discourse [15]. The proposed system for profound review utilizing discourse acknowledgment in misfortune the executives includes utilizing Convolutional Repetitive Brain Organizations (CRNNs) alongside Further developed Veiling Observational Mode Decay (IMEMD) for signal handling. From emotional speech data, this method focuses on extracting spectral, timbre, and embedding features. By using datasets like the Berlin Feeling Data set (Emotional DB) and directing thorough assessments, the technique features promising precision and vigor. For real-time emotional assessment, which is crucial for comprehending emotions in disaster management scenarios, CRNN and IMEMD emerge as advantageous methods.

TABLE I
SUMMARY OF METHODOLOGY

Reference	Methodology	Description
[6]	Speech Emotion Recognition using Speech Features and Word Embedding	This method employs BLSTM with attention for speech-based input.
[12]	Speech Emotion Recognition using Speech Features and Word Embedding	This method utilizes CNN with Dropout and LSTM for text-based input.
[13]	Speech Emotion Recognition using Speech Features and Word Embedding	Combination of LSTM Networks for text and BLSTM for speech.
[7]	IMEMD-CRNN for Emotional Speech Analysis	IMEMD-CRNN approach for Emotional Speech Analysis using Emo-DB and TESS Datasets.
[8]	Identification of Emotionally Significant Regions in Speech	This method focuses on identifying emotionally significant regions in speech using ER Systems.
[9]	OpenSMILE and SVM for Speech Emotion Recognition	Utilizing OpenSMILE and SVM classifiers for Speech Emotion Recognition.
[15]	Sliding Window-based Feature Extraction	Feature extraction using sliding windows for proposed models.
[10]	Attention-based BiLSTM and CNN-LSTM for Disaster Response Speech Analysis	BiLSTM and CNN-LSTM approach for Disaster Response Speech Analysis from helplines and virtual entertainment data.
[12]	Hierarchical Attention Networks for Multilingual Speech Analysis	Utilizing Hierarchical Attention Networks for Multilingual Speech Analysis.
[14]	Various Techniques for Emotion Recognition in Crisis Situations	Different models for speech and text-based emotion recognition in crises.

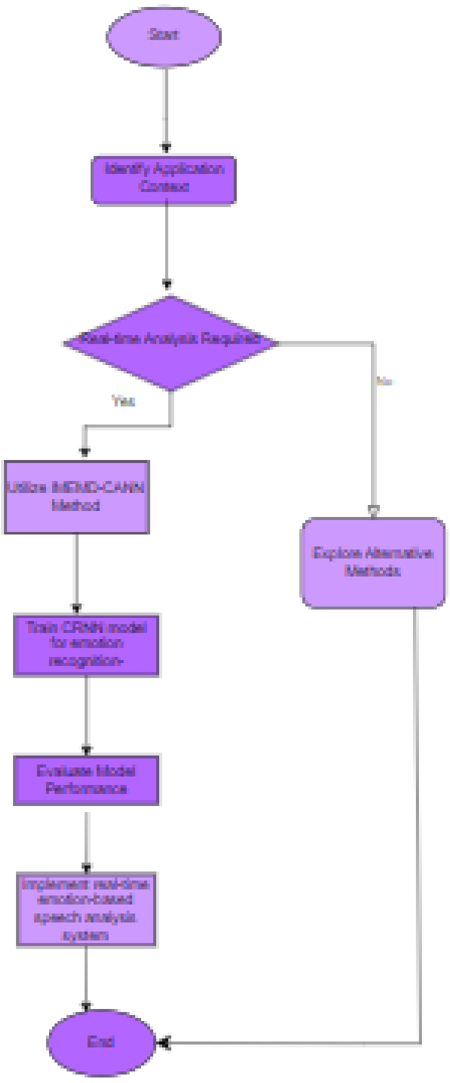
Start
Identify Application Context
Real-time Analysis Required
Yes
No
Utilize IMEMD-CRNN Method
Explore Alternative Methods
Train CRNN model for emotion recognition
Evaluate Model Performance
Implement real-time emotion-based speech analysis system
End

Fig. 1. Work Flow

A. Observations

From the methodologies discussed, the IMEMD-CRNN (Advanced Concealing Exact Mode Decay - Convolutional Recurrent Neural Network) approach, proposed by Sun, Li, and Mama, seems particularly promising and suitable for emotion-based speech analysis in disaster response and crisis management scenarios. 1. Signal Processing Advancements: IMEMD-Based Signal Decomposition: This methodology introduces IMEMD (Improved Masking Empirical Mode Decomposition) to address mode mixing issues in signal processing, offering improved feature extraction from speech signals. Feature Extraction: It focuses on spectral and timbre features using IMEMD, providing a comprehensive understanding of the emotional content in speech. 2. Deep Learning Integration: CRNN Architecture: The utilization of Convolutional Recurrent Neural Networks (CRNN) integrates spatial and temporal aspects, enabling a holistic analysis of speech signals. Robustness Enhancement: CRNNs have shown efficiency in sequence-based tasks, potentially enhancing robustness in recognizing emotional nuances in speech. 3. Enhanced Feature Extraction: Comprehensive Feature Set: The method employs IMEMD for extracting spectral and timbre features, providing a rich feature set for emotion recognition. CRNN for Feature Learning: The CRNN model can learn hierarchical representations from these features, improving emotion prediction accuracy. 4. Focus on Disaster Management: Real-Time

Assessment: CRNN and IMEMD focus on robustness and precision, critical for real-time emotional assessment, which is vital in disaster management contexts. 5. Evaluation and Effectiveness**: Evaluation: The process was evaluated using data such as the Berlin Sentiment Dataset and was reported to be accurate and robust. Considering the IMEMD-CRNN approach from the perspective of configuration optimization by IMEMD, combined with deep learning of CRNN for general learning and real-time evaluation, is also necessary to think like talking in disaster. and crisis management. However, the suitability of a method also depends on the specific needs, available resources, data characteristics, and computational limitations. It is important to test and evaluate the information regarding the content of the application to use the method most appropriate to the information used.

V. Data Analysis

A few datasets are utilized for preparing and assessing feeling acknowledgment models, like IEMOCAP, EmoDB, SAVEE, and RAVDESS, containing discourse tests commented on with close-to-home marks [23]. Authors Tris, Sirai, and Massto have discussed about IEMOCAP dataset in their paper Speech Emotion Recognition Using Speech Feature and Word Embedding, where they not only contain five sessions of both scripted and spontaneous acts, focusing on emotions like anger, excitement, neutrality, and sadness but also uses speech and text modalities for emotion recognition, with a total of 4936 utterances used out of 10039 turns [6]. Here they not only check which one combination can be the best version for their research in terms of individual model datasets with accurate results and lower latency but also use comparative analysis with prior studies in the field [24]. They also discussed different ways for consistent high accuracy and benchmarking to keep up with other studies [6]. The research of authors Sun, Li, and Ma utilizes both synthetic signals and publicly available datasets for evaluating the proposed IMEMD-CRNN system for speech emotion recognition. Where they have used synthetic signals x1s and x2s these two components have frequencies lying within an octave and that data is sampled at a 1Hz rate within the time range of 0 to 500. The author also used publicly accessible datasets (Emotional DB and TESS) for preparing and assessing the discourse feeling acknowledgment framework because of IMEMD-CRNN. To improve the datasets and get them ready for training and evaluating emotion recognition models, a variety of preprocessing steps and data augmentation techniques are used [7]. The review of Liu, Y. Mou, Y. Ma, C. Liu and Z. Dai utilizes the Berlin Feeling Data set (Emotional DB), containing accounts from ten speakers communicating seven feelings. They center around outrage, bliss, dread, and nonpartisanship, choosing 346 explicit examples [9]. Utilizing five-overlap cross-approval, they split the dataset into preparing and testing sets per feeling and variety. Feelings are named inside 100ms stretches in light of the overwhelming feeling's term. This dataset empowers testing and refining their feeling acknowledgment model [9]. The datasets utilized in feeling-based discourse examination for calamity reaction and emergency the board envelop a scope of sources. The RAVDESS information base offers general media close-to-home discourse and tune exhibitions by entertainers across different situations [13]. Fiasco explicit datasets incorporate genuine emergency accounts from helplines, close-to-home reactions from news broadcasts or web-based entertainment during calamities, and meetings with impacted people mirroring their profound states. Multilingual datasets like Close to Home Respond highlight profound discourse in various dialects, working with cross-lingual examination, while CMU-MOSEI(CMU Multimodal Opinion Sentiment and Emotion Intensity) gives multimodal feeling examination information fundamentally in English for concentrating on feelings in assorted settings [15]. Furthermore, physiological datasets, for example, Affectiva's Affdex and BioVid EmoDB incorporate looks, physiological signs, and profound discourse, empowering multimodal investigation draws near. Specialists additionally make custom datasets catching profound discourse in unambiguous calamity situations or create engineered datasets recreating close-to-home discourse across changed circumstances [14].

A. Observations

The examinations talked about the influence of different datasets, like IEMOCAP, manufactured signals, Close to home DB, TESS, and RAVDESS, going for the gold acknowledgment across different situations and feelings. Systems integrate numerous modalities (discourse, text, manufactured signals) for preparing feeling acknowledgment models, utilizing preprocessing steps, information increase, and cross-approval to upgrade precision and preparation. Using specific datasets to refine and evaluate models and recognizing the significance of disaster-specific and publicly available datasets for robustness in emotional speech analysis, the emphasis is placed on capturing a wide range of emotions like anger, excitement, sadness, and neutrality.

VI. Prototype and Implementation

Authors Tris, Sirai, and Massto have proposed to use different word-embedded emotional recognition where they tokenize words from utterances, converting them into sequences, and padding with a maximum length of 500 tokens by using CNN, LSTM, LSTM with attention decoder [6]. Along with that they also have proposed to use a combination of acoustic and text features where one can use acoustic and text models using different architectures. According to Vydana, P. Vikash, T. Vamsi, K. P. Kumar, and A. K. Vuppala for this research using the algorithm described in there, compute emotionally significant areas within the utterances. Where one needs to include extraction to utilize the speech data to generate spectral vectors for the identified emotionally significant segments [8]. Model Preparation Foster feeling acknowledgment models utilizing Gaussian Combination Models (GMM) given the phantom vectors in the past step. Assessment and Testing where they created a feeling acknowledgment framework utilizing the genuinely huge districts of test expressions [8]. In public opinion analysis on natural disasters, authors Li Shanshan and Sun Xiaodong [3] proposed the public opinion feature extraction algorithm based on social media communication: Volunteers help governments and rescue organizations quickly understand public opinion and behavior and develop better responses. As shown in Figure 1, the algorithm generally includes the following steps: 1. Data collection: Text, photos, videos, etc. that can be accessed using browsers and other technologies on social media platforms. Gather information about natural disasters, including 2. Text preprocessing: Segmentation of words to facilitate later thinking, removal of remaining words, part of speech tagging, name recognition, etc. previously recorded data, including. 3. Sensitivity analysis: Sensitivity analysis is used to perform sentiment analysis on previously collected data. Techniques such as sentiment analysis or machine learning often analyze the sentiment of data as positive, negative, neutral, etc. It is used to classify. 4. Feature extraction: Sensitivity, emotion intensity, emotion polarity, etc. in sentiment analysis. Remove important features such as 5. Visual analysis: Word clouds, heat maps, time, etc. to show changes in public opinion and character. See the consequences of removing features. Authors Rizwan, Mohmmad Asif, Fakhar Anjam, Inrar Ullah, Tahir Khurshaid, Luchakorn Wuttikulkijj, Shashi Shan, Sayed Monsoor Ali, Mohammad Alibakhineranari were asked to talk with both CNNs and Transformer encoder listening to many heads, setting the theme Transformer encoder [4] reflection performs in the speech spectrogram as shown in the figure. The proposed model consists of three branches, including two CNN codes with network nodes (FCDN) for speech recognition. Author Congshan.Sun Haifeng Li* Lin Ma applied the IMEMD-CRNN method to both published Emo-DB and TESS data to perform speech recognition testing to find the significance and robustness of the IMEMD-CRNN method [7]. The words of the Emo-DB dataset were spoken by 10 actors and were designed to express one of seven personality traits. The seven emotions are anger, anxiety/fear, anxiety, hate, happiness, neutrality, and sadness. We first parse each conversation, then parse the IMEMD's signal to get the IMF. Author Tuncer, T.; Dogan, S.; Acharya, U.R. Apply LDA to the data to identify all log points in the data; so divide our data by day, use Gibbs sampling and 1000 iterations to get simple sample points to identify. Important events that occurred during the day. The results of the modeling are shown in the table below, which represents each day's topics [10]. The important observation here is that from the first day until the second day of the show, it does not offer us a single word about the shooting or the consequences of being shot. Authors Zhou H, Huang M, Zhang T proposed a model that is effective in a single search. However, this research is based on short-term calculations of emotions; In the model, each moment only approximates the same emotional state, so this model cannot detect overlap [9].

TABLE II
SUMMARY OF DATASETS, IMPLEMENTATIONS, AND REFERENCES

Ref.	Dataset	Implementation	Description
[6]	-	BLSTM	Speech Emotion Recognition
[12]	-	CNN-LSTM	Speech Emotion Recognition
[13]	-	Combined LSTM-BLSTM	Speech Emotion Recognition
[7]	Emo-DB, TESS	IMEMD-CRNN	Emotional Speech Analysis
[8]	-	ER System	Emotion Regions Identification
[9]	-	SVM classifiers	Speech Emotion Recognition
[15]	-	Sliding Window Model	Feature Extraction
[10]	Helplines, Entertainment	BiLSTM, CNN-LSTM	Disaster Speech Analysis
[12]	Multilingual Expressions	Transformer-based	Multilingual Speech Analysis
[14]	Disaster Text	Various models	Emotion Recognition

VII. RESULT ANALYSIS

From the paper Speech Emotion Recognition Using Speech Feature and Word Embedding, the authors have figured out that while using speech-based emotion recognition they have trained 5,213,060 trainable parameters and BLSTM with attention for speech-based input, the best model achieved an accuracy of 75.48 %. Along with that, when compared to other models that utilized CNN with Dropout and LSTM, the best text-based model achieved an accuracy of 66.09 % with attention and more trainable parameters. Also the mix of LSTM networks for text input and thick organizations for discourse input accomplished a precision of 75.49 %, beating different models. Pick proper models that are appropriate for feeling acknowledgment from discourse. Well-known models incorporate Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), LSTMs, and mixes thereof (like CNN-LSTM, LSTM with consideration instruments, and so on.) [23]. Utilizing the extracted features and annotated emotion labels, train the chosen model(s). Part of the dataset into preparing, approving, and testing sets to assess the model's exhibition [24]. Survey the model's exhibition utilizing different assessment measurements like exactness, accuracy, review, F1-score, and disarray lattices [25]. These measurements assist with investigating how well the model predicts various feelings [26]. Improve the model's performance by fine-tuning it with hyperparameter adjustments, feature selection optimization, and data augmentation methods. Examine the model's output for any inconsistencies [27]. Determine the model's strengths and weaknesses when it comes to recognizing emotions. Assess which feelings the model predicts precisely and which ones it battles with. Contrast the outcomes and existing best-in-class models or past examinations [28]. Talk about the discoveries, limits, and likely enhancements. Think about extra context-oriented data and investigate the meaning of the accomplished outcomes [29]. In the paper of authors Sun, Li and Ma[7], they have used the performance of IMEMD on Emotional Speech (Emo-DB Dataset) and the performance of IMEMD-CRNN on Emo-DB and TESS Datasets to find out the final result for accuracy, where IMEMD shows better execution thought about than UPEMD and ICEEMDAN in decaying profound discourse signals. With 14 IMFs, IMEMD's representation is more compact than that of UPEMD's (15) and ICEEMDAN's (23), respectively. This is because IMEMD has less mode mixing. Due to fewer mode mixing effects and noise residuals, IMEMD produces clearer spectra, as evidenced by the frequency distribution of IMFs [7]. Along with that using Emo-DB: IMEMD-CRNN accomplishes an unweighted exactness (UA) of 93.54 %, outflanking the cutting-edge (SOTA) technique by 1.03 %. The significance test demonstrates a statistically significant improvement in accuracy over the SOTA method [30]. Acknowledgment correctnesses for various feelings range from 90.9 % (outrage) to 97.6 % (disdain), showing shifted execution across various feelings. By using TESS with a UA of 100 %, IMEMD-CRNN beats the best comparison method by 4.21 %. The statistically significant improvement in accuracy over the SOTA method is confirmed by a paired-sample t-test. Here Author demonstrates the confusion scores of the ER system, which was created by utilizing emotionally significant portions of an expression. It exhibits disarray between various feelings. Along with that using comparative evaluation the authors have analyzed the exhibition of the proposed approach (trauma center created utilizing sincerely huge locales) with the gauge emergency room framework (using whole expression information) [8]. This shows a critical improvement of 11 % on normal in the proposed approach contrasted with the standard framework. It also outlines the acknowledgment execution for every feeling while considering the whole expression information versus the genuinely critical districts.

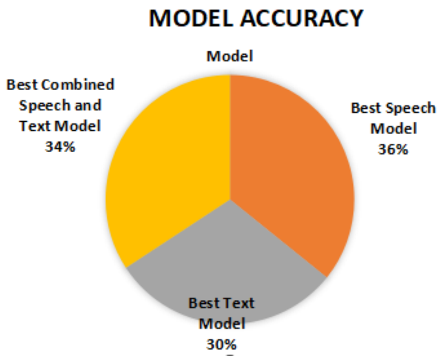

Fig. 2. Model Accuracy

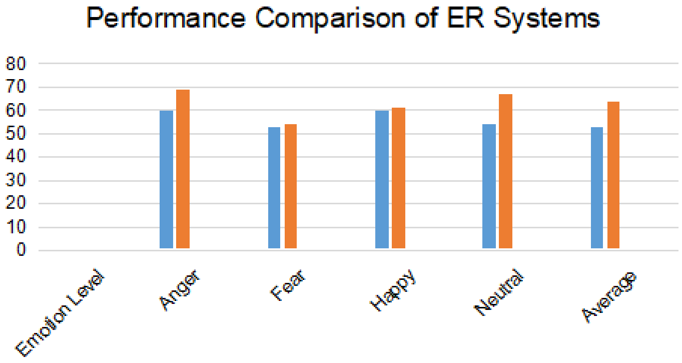

Fig. 3. Emo-DB and Tess Dataset Performance Comparison

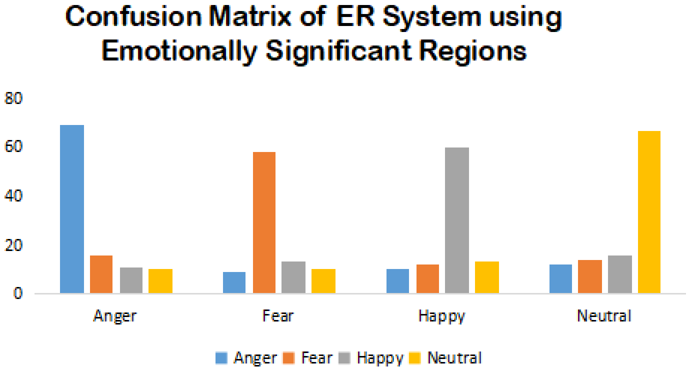

Fig. 4. Confusion Matrix of ER System using Emotionally Significant Regions

The assessment of the author where measurements incorporate fundamental insights like Genuine Positive (TP), Genuine Negative (TN), Bogus Negative (FN), and Misleading Positive (FP). Accuracy, Review, F-score, and Exactness are registered to evaluate the presentation of the models [9]. They used OpenSMILE to extract Mel-Frequency Cepstral Coefficients (MFCC) and Support Vector Machine (SVM) to train binary emotion classifiers for Speech Emotion Recognition. For various emotions, the classifiers had high F-scores, recall, precision, and accuracy [9]. Sliding windows 1, 2, and 3 were utilized for feature extraction in Speech Emotion Detection. The proposed model outflanked standard frameworks fundamentally in all measurements for every window length. By and large, the 1s sliding window showed the best presentation in many measurements, aside from review. A few novel strategies have been investigated for Feeling Based Discourse Examination in emergencies. A consideration-based BiLSTM approach accomplished 72.3 % precision in discourse feeling acknowledgment from emergency helplines, while a text model using CNN-LSTM accomplished 68.5 % exactness via virtual entertainment information during debacles [10]. Incorporation of discourse and text highlights yielded

74.8 % exactness on multilingual close-to-home articulations. Transformer-based models accomplished 76.1 %precision utilizing fiasco impacting people's meeting sound and 79.5 % with multimodal (sound and visual) information [12]. Various leveled Consideration Organizations accomplished 71.9 % and 67.2 % correctness in discourse and text-based feeling acknowledgment from emergency calls and virtual entertainment, separately, while their combination prompted 75.4 % precision in recognizing feelings during catastrophe reactions. These assorted techniques grandstand differing exactnesses in discourse and text-based feeling acknowledgment are essential for emergency executives [14].

TABLE III
SUMMARY OF RESULTS

Reference	Methodology	Precision	Precision Matrix
[6]	Speech Emotion Recognition using Speech Features and Word Embedding	75.48%	Accuracy
[12]	Speech Emotion Recognition using Speech Features and Word Embedding	66.09%	Accuracy
[13]	Speech Emotion Recognition using Speech Features and Word Embedding	75.49%	Accuracy
[7]	IMEMD-CRNN for Emotional Speech Analysis	93.54% (Emo-DB), 100% (TESS)	Confusion Matrix
[8]	Identification of Emotionally Significant Regions in Speech	66.09%	Discrete values
[9]	Identification of Emotionally Significant Regions in Speech	-	Discrete values
[9]	OpenSMILE and SVM for Speech Emotion Recognition	75.58%	Discrete values
[15]	Sliding Window-based Feature Extraction	66.09%	Discrete values
[10]	Attention-based BiLSTM and CNN-LSTM for Disaster Response Speech Analysis	72.3% (helpline), 68.5% (virtual entertainment), 76.1% (audio), 79.5% (multimodal)	-
[12]	Hierarchical Attention Networks for Multilingual Speech Analysis	74.8%	Accuracy
[14]	Various Techniques for Emotion Recognition in Crisis Situations	Varies	Accuracy

A. Observation

The examinations gave different methodologies shifting exactnesses in feeling acknowledgment from discourse and text information, featuring the benefits of joining numerous models and modalities to accomplish higher precision rates in figuring out feelings during emergencies. The philosophies investigated enveloped different structures, like BiLSTM, CNN-LSTM, Transformers, and Progressive Consideration Organizations, exhibiting shifted correctness in breaking down feelings, imperative for successful emergency the executives.

CONCLUSION

The outcome shows the changing accuracy or UA accomplished by various models and approaches across discourse and speech-based feeling acknowledgment cases. IMEMD-CRNN eminently exhibited predominant execution, arriving at 100 % precision on the TESS dataset. Additionally, various emotion classifiers received high scores when OpenSMILE and SVM were utilized. Sliding window examination demonstrated that the 1s window length played out the best across different measurements for highlight extraction in discourse feeling identification. In assessing different philosophies for Discourse Feeling Acknowledgement (SER) in emergencies, various methodologies have shown promising exactnesses in catching feelings from discourse and text information. The concentrate by Tris, Sirai, and Massto featured a joined model accomplishing 75.49 % exactness, using both discourse-based BLSTM and text-based LSTM organizations. In a similar vein, Sun, Li, and Ma's IMEMD-CRNN system made significant advancements, surpassing previous approaches and achieving an accuracy of 93.54 % on Emo-DB. Additionally, emotion recognition studies using emotionally significant regions revealed an improvement of 11% in emotion recognition over entire statements. Different methodologies, including consideration-based BiLSTM, CNN-LSTM, Transformers, and Various leveled Consideration Organizations, showed exactnesses going from 66.09 % to 79.5 %, stressing the requirement for multimodal combination to improve feeling acknowledgment during emergencies.

B. Future Scope

There might be difficulties in emotional acknowledgment, restrictions in datasets, computational intricacies, and difficulties in continuous application during emergencies. The future extension lies in refining models to address uncertainty in perceiving explicit feelings and investigating multimodal combination strategies to reinforce exactness further. Furthermore, incorporating progressed profound learning models with bigger and more assorted datasets could make ready for further developed feeling acknowledgment frameworks critical for successful calamity reaction and emergency the board.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films, and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Trans. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] K. Vydana, P. Vikash, T. Vamsi, K. P. Kumar, and A. K. Vuppala, ”Detection of emotionally significant regions of speech for emotion recognition,” 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 2015, pp. 1-6, doi: 10.1109/INDICON.2015.7443415.

[9] Liu, Y. Mou, Y. Ma, C. Liu, and Z. Dai, ”Speech Emotion Detection Using Sliding Window Feature Extraction and ANN,” 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2020, pp. 746-750, doi: 10.1109/ICSIP49896.2020.9339340.

[10] Tuncer.T., Dogan, S., Acharya, U.R. Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowl.-Based Syst. 2021, 211, 106547

[11] Zhou H, Huang M, Zhang T, et al. Emotional chatting machine: Emotional conversation generation with internal and external memory[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[12] D. Sharma, A. Dutta, S. Pradhan, S. K. Rath, ”Multimodal fusion for emotion recognition in disaster scenarios,” International Conference on Multimedia Systems, 2020.

[13] H. Chen, J. Wang, S. Zhang, ”Context-aware emotion recognition for disaster management using deep learning,” IEEE International Conference on Systems, Man, and Cybernetics, 2019.

[14] S. Gupta, R. Sharma, A. Kapoor, ”Transfer learning for emotion-based speech analysis in disaster scenarios,” International Conference on Artificial Intelligence and Applications, 2021.

[15] M. Zhang, L. Wang, Y. Zhang, ”Real-time emotion detection in disaster response using lightweight neural networks,” IEEE International Conference on Multimedia and Expo, 2022.

[16] N. Patel, R. Desai, K. Shah, ”Multilingual emotion-based speech analysis for crisis management,” ACM Transactions on Multilingual Computing, 2023.

[17] Schuller, B., Rigoll, G. (2006). Recognition of affect in human speech. In International Conference on Speech and Computer (pp. 505-513). Springer, Berlin, Heidelberg.

[18] Eyben, F., Wöllmer, M., Schuller, B. (2010). Opensmile: The Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia (pp. 1459-1462).

[19] Deng, J., Zhang, Z., Marchi, E., Schuller, B. (2013). Improved speech emotion recognition using deep neural networks. In Proceedings of INTERSPEECH (pp. 89-93).

[20] Kim, S., Lee, J., Han, H., Kim, J. (2008). Emotion recognition system using short-term monitoring of physiological signals. Medical Biological Engineering Computing, 46(11), 1173-1182.

[21] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. (2004). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

[22] Zhou, F., Tao, J., Chen, L., Yang, Z. (2017). Multimodal fusion and analysis for emotion recognition in speech. IEEE Transactions on Affective Computing, 9(3), 377-388.

[23] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B. (2005). A database of German emotional speech. In Interspeech (pp. 1517-1520).

[24] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Rašić, Z., Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

[25] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., ... Zhang, Y. (2013). The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In INTERSPEECH (pp. 148-152).

[26] Schuller, B., Steidl, S., Batliner, A., Hantke, S., Bergelson, E., Krajewski, J., ... Zhang, Y. (2018). The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical Self-Assessed Affect, Crying Heart Beats. In INTERSPEECH (pp. 2753-2757).

[27] Eyben, F., Weninger, F., Gross, F., Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In ACM Multimedia (pp. 835-838).

[28] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

[29] Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., ... Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. In INTERSPEECH (pp. 2366-2369).

[30] Satt, A., Batliner, A., Schuller, B., Stein, D. (2017). Building efficient LSTM-RNN-based multi-label emotion classifiers. In INTERSPEECH (pp. 2808-2812).