Diverse machine learning models for Speech Emotion Recognition.

Sonu P
Department of Computer Science and Applications
Amrita School of Computing
Amrita vishwa Vidyapeetham
Amritapuri, India
sonu@am.amrita.edu

Vaishnav Babu
Department of Computer Science and Applications
Amrita School of Computing
Amrita vishwa Vidyapeetham
Amritapuri, India
vaishnavbabu1420@gmail.com

Anuj M
Department of Computer Science and Applications
Amrita School of Computing
Amrita vishwa Vidyapeetham
Amritapuri, India
anujms2k@gmail.com

***Abstract*—This work gives a thorough investigation into voice emotion recognition, with a focus on the comparison of machine learning models for this task. It explores the classification of speech data into various emotions, regardless of semantic content, including emotions like happiness, sadness, anger, and neutrality. The primary objective of SER systems is to efficiently detect emotions and incorporate human elements, such as emotional response, into machines. The study investigates and compares several popular machine learning algorithms, including support vector machines (SVM), Classification And Regression Tree (CARTS) , Long short-term memory (LSTM) .By enabling machines or robots to recognize and respond to emotions, SER systems reduce human time and effort while creating a more empathetic and natural interaction. Overall, this research contributes to the advancement of speech emotion recognition techniques and provides a comparative analysis of machine learning models, enabling researchers and practitioners to make informed decisions when designing and implementing emotion-aware systems.**

***Index Terms*—Speech Emotion Recognition, SVM, LSTM, CARTS , TESS .**

I. Introduction

Speech emotion recognition (SER) is a crucial technology that addresses the challenge of detecting emotions in speech. Its primary purpose is to classify speech data into different emotional categories, regardless of the semantic content conveyed. Emotions such as happiness, sadness, anger, and neutrality are commonly targeted by SER systems. The significance of SER lies in its ability to efficiently detect emotions and infuse human elements, such as emotional response, into machines. By recognizing and understanding emotions in speech, SER systems enable machines to engage in meaningful dialogue with humans, fostering more natural and empathetic interactions. This has wide-ranging implications, including applications in fields like mobile phone use and autonomous driving. Implementing well-working SER systems not only reduces human time and effort involved in emotion recognition tasks but also enhances human-machine communication by allowing machines or robots to respond appropriately and create a more meaningful connection. Overall, SER systems are becoming increasingly important as they bridge the gap between humans and machines, bringing a human-like element to machine interactions and improving overall user experience.

Speech emotion recognition is a captivating field of study that has gained significant attention in recent years. The ability to detect and analyze emotions expressed through speech holds immense potential in various applications, such as human-computer interaction, sentiment analysis, and mental health monitoring. In this research paper, we delve into the fascinating world of speech emotion recognition, exploring its importance and discussing the current state of research. Furthermore, we conduct a comprehensive comparison of different machine learning models employed in this domain, highlighting their strengths, weaknesses, and performance. Understanding and interpreting emotions is a fundamental aspect of human communication. Traditionally, emotion recognition heavily relied on visual cues, such as facial expressions and body language. However, speech offers a unique opportunity to extract emotional information, as it contains rich and complex cues that reflect an individual's internal state. By analyzing the acoustic and linguistic features present in speech signals, we can unravel the emotional content and gain valuable insights into human affective states.

When examining its prospective uses, the relevance of speech emotion recognition becomes clear. Emotion-aware systems in human-computer interaction can change their replies based on the user's emotional state, resulting in more personalized and engaging experiences. Understanding user emotions, for example, can permit appropriate responses and boost user happiness in virtual assistants or chatbots. Furthermore, accurate speech emotion identification can improve sentiment analysis in areas such as social media monitoring and customer feedback analysis, allowing organizations to measure public opinion and make informed decisions appropriately.

II. LITERATURE SURVEY

Human-computer interaction (HCI) research on automatic speech emotion recognition (SER) has a wide range of applications [1]. To identify emotions including anger, happiness, sorrow, neutral, and fear, this work uses the Berlin Emotional Database and SVM as a classifier. The system extracts features like MFCC and MEDC from the database and achieves high accuracy in gender-independent, male, and female speech. The research highlights the importance of

feature extraction and evaluates different kernel functions. Results indicate SVM's accuracy of 93.75% for gender-independent cases, 94.73% for male speech, and 100% for female speech. The paper concludes by discussing potential applications of SER and the effectiveness of SVM and speech features in emotion recognition.

The paper [2] examines gender differences in vocal apparatus and expressive emotional responses. Women tend to express more intense disgust and horror, while men show stronger anger. The study evaluates results using confusion matrices for 16 and 8 emotion classes with gender abstraction and grouping. LSTM achieves 72% accuracy for 16 classes, while CNN-LSTM achieves 94%. When considering separate male and female speakers, LSTM improves by 6% for both genders. CNN-LSTM performs well for mixed gender and gender-separated classifications, with faster training in 20 epochs. However, the neutral class detection accuracy is only 50%, and confusion rates are higher for low arousal emotions.

Deep learning techniques, such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM), have advantages over traditional methods [3]. They can detect complex structures and features without manual feature extraction, handle unlabelled data, and extract low-level features from raw data. DNNs and CNNs are efficient for image and video processing, while RNNs and LSTM are effective for speech-based classification and natural language processing (NLP).

The SVM classifier uses a binary decision technique and bases its classification solely on the MFFC feature [4]. The CART classifier uses a binary recursive decision process, and it depends on all 13 retrieved features for classification. The outcomes of these two algorithms are also contrasted. The recognition rate reported by the SVM is 92.01. When opposed to SVM, CART performs worse at distinguishing emotions when the training set's percentile is lower. With a recognition rate of 91.78%, CART likewise provides a rate that is closer to the SVM as the training set percentile is raised. To make searching easier, the CART algorithm adds more groups to the file's inner clusters. CART also places a lot of focus on searching within the inner cluster.

The work presents a novel approach for decision trees and random forests for vocal-based emotion identification [5]. The technique categorizes speech signals according to many emotion types, such as surprise, disgust, neutral, fear, and happiness. Pitch, intensity, formants, autocorrelation, and noise-to-harmonics ratio are a few of the features that are recovered from the speech signals. Experimental findings on the SAVEE Database show an average recognition rate of 66.28 percent, with happiness receiving the highest rate of 78 percent. On the same datasets, the suggested technique performs better than deep neural networks and linear discriminant analysis, demonstrating its effectiveness in recognizing vocal emotions. The work emphasizes potential uses for emotion-based disease detection and human-robot interaction.

This paper [6] focuses on speech emotion analysis and its potential application in affective computing. The authors propose new features based on harmonics and Zipf's law to better characterize speech emotions in terms of timbre, rhythm, and prosody. They also introduce a multi-stage classification scheme based on a dimensional emotion model for improved emotional class discrimination. The effectiveness of their approach is demonstrated using the Berlin dataset and the DES dataset, achieving classification rates of 68.60% and 81%, respectively. The paper addresses the challenge of bridging the semantic gap between low-level speech signals and high-level emotional information and discusses previous research in automatic speech emotion recognition. Overall, it presents a novel approach with promising results in the field.

This work 13 creates a speech emotion recognition system for the Arabic-speaking community, overcoming a lack of emotional speech databases in many languages. The researchers develop a database of emotions elicited from 14 non-native Arabic speakers, classifying them using various approaches and supervised learning algorithms such as SVM and ELM. They assess system performance using accuracy, specificity, precision, and recall, and they investigate multistage classification systems that improve recognition. They illustrate the efficiency of their approach for Arabic speech emotion recognition by comparing their results to the Emo-DB database, a typical emotional speech corpus. This research contributes to affective computing by filling a gap in Arabic emotional speech databases and offering enhanced classification techniques.

This paper [14] emphasizes the significance of automatically recognizing emotions in speech and the challenges in evaluating emotion recognition engines. It highlights the limitations of using acted emotions for evaluation and the need for cross-corpus evaluations to assess system generalization. The passage mentions several language resources and databases used in emotion recognition research. It discusses the normalization strategies applied and the classification technique used, which involves supra-segmental feature analysis and Support Vector Machines (SVM). The results demonstrate the impact of training and test sets, normalization methods, and the limitations of current systems. Future directions include addressing language and cultural differences and improving feature selection and adaptation strategies.

An open-source feature extraction toolbox called openSMILE integrates approaches from the music information retrieval and voice processing communities [15]. It supports a number of low-level audio descriptors, including loudness,

Mel-frequency cepstral coefficients, linear predictive coefficients, CHROMA and CENS features, and more. Low-level descriptors can be subjected to statistical functionals and delta regression. The toolkit is quick and compatible with Unix and Windows platforms because it is implemented in C++ without any third-party dependencies. Its modular architecture makes straightforward extension through plug-ins possible. Along with batch and off-line processing, openSMILE also allows online incremental processing.According to performance benchmarks, openSMILE operates effectively in real-time with minimal real-time overhead for feature extraction activities. It has proven effective in research initiatives and problems involving emotion recognition, paralinguistic analysis, and speaker classification.

This study [16] compares two popular approaches for recognizing emotions in speech: suprasegmental modeling with systematic feature brute-forcing and frame-level modeling with Hidden Markov Models (HMM). Nine common emotion corpora are compared in the study.The corpora contain a wide range of emotional expressions, including more archetypal occurrences as well as spontaneous speech. For easier comparison, the emotions in each database are grouped into binary valence and arousal discrimination tasks. The results reveal large variations between the corpora, and supra-segmental modeling is advantageous when tackling several classes at once. A description of each corpus' features is also included in the abstract.

This research paper [17] used various acoustic features and Gaussian mixture models (GMMs) to automatically detect emotions in speech. The study evaluated the performance of MFCCs, MFCC-low (which models pitch), and pitch features using GMMs on the frame level. Results showed that combining the three classifiers significantly improved the performance for emotion classification. The study acknowledges the challenges in labeling spontaneous emotions and aims to enhance human-machine interaction systems. The research was conducted on two different corpora: Swedish voice-controlled telephone services and English meetings. The study suggests future directions for improving emotion detection, such as considering dialogue-level analysis and incorporating lexical content. The research was supported by the European Commission's 6th Framework Program.

The representation of emotions and affect is a crucial decision in developing an automatic affect analyzer [18]. There are two main approaches: (1) the social sciences approach, which is based on basic emotion categories recognized universally, and (2) the dimensional approach, which represents affective states as points in a dimensional space. The social sciences approach is commonly used and categorizes affective displays into discrete classes, while the dimensional approach captures the relationships and transitions between affect dimensions. The most widely used dimensional models are the Circumplex of Affect and the Pleasure-Arousal-Dominance (PAD) emotion space, which represent affect along arousal, valence, and sometimes power dimensions.
This paper [19] describes a new model for continuous emotion identification from speech that employs deep learning techniques. The proposed model consists of a Convolutional Neural Network (CNN) for feature extraction and a 2-layer Long Short-Term Memory (LSTM) network to capture contextual information. The model is trained end-to-end and outperforms state-of-the-art methods in terms of concordance correlation coefficient on the RECOLA database. The paper discusses the challenges of emotion recognition, the use of deep neural networks in speech emotion recognition, and the design of the proposed model. The authors also compare their results with previous studies and demonstrate the effectiveness of their approach. The experiments are conducted on the RECOLA dataset, and the results show improved performance in both arousal and valence dimensions. The paper concludes by suggesting future work on deeper CNN models and larger databases for audio analysis.
This research [20] presents an overview of deep learning algorithms for speech-based emotion detection (SER) in human-computer interaction (HCI). It goes through the benefits of deep learning for extracting complicated features from speech signals and dealing with unlabeled data. Various deep learning models are examined, including DNNs, CNNs, RNNs, DBMs, and DBNs. The report also investigates various forms of emotion databases utilized in SER research. The benefits and drawbacks of deep learning in SER are discussed, and future research directions are suggested. Overall, this work is a wonderful resource for field researchers and practitioners.

III. Methodology

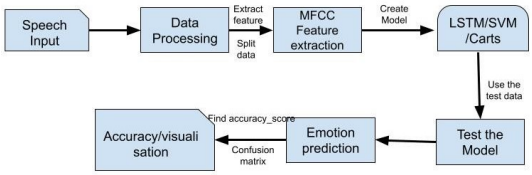

Fig. 1. Architecture Diagram

A. Speech Input:

The speech input can be in the form of audio recordings or real-time audio streams captured from a microphone or any other audio source. These audio samples contain spoken words or vocal expressions that convey emotional information.

The speech input is passed through the preprocessing stage, where it undergoes various steps, including noise removal, segmentation, and framing. These steps prepare the speech signal for further analysis and feature extraction.

The system advances to the feature extraction stage after the speech signal has been preprocessed. Relevant acoustic information from the speech signal, such as MFCCs, pitch,

energy, and spectral properties, are retrieved here. These qualities provide crucial information on the vocal characteristics and dynamics of the speech.

B. Data Processing:

A digital converter converts the incoming spoken data into digital signals during this step. Following that, the signal is processed through a band-pass filter to remove low and high frequency noise. Data processing is a crucial step in speech emotion recognition, where the raw speech data is transformed and prepared for analysis by extracting relevant features. The process involves several key steps:

- Noise Removal: Background noise, such as ambient sounds or recording artifacts, is removed to enhance the quality of the speech signal. Techniques like spectral subtraction or adaptive filtering can be employed.
- Segmentation: To identify individual utterances or meaningful speech units, the continuous speech stream is split into smaller parts. Segmentation can be performed based on silence detection, fixed time intervals, or using voice activity detection (VAD) algorithms.
- Framing: Each segmented speech segment is further divided into shorter frames of fixed duration (e.g., 20-30 milliseconds) to capture temporal variations in speech.

C. Feature Extraction:

Mel-frequency Cepstral Coefficients (MFCCs): MFCCs are commonly used characteristics in emotion recognition in speech. They extract the log-scaled Mel filterbank energies and perform a discrete cosine transform (DCT) to generate the cepstral coefficients to represent the spectral envelope of speech.

- Pitch: Pitch or fundamental frequency (F0) represents the perceived pitch of the speaker's voice. It provides information about intonation and prosody and can be estimated using techniques like autocorrelation or cepstral analysis.
- Energy: The overall intensity or loudness of the spoken signal is represented by energy. It is frequently estimated as the squared magnitude of the speech waveform or in frequency subbands.
- Spectral Features: The distribution and dynamics of the spectrum content in distinct frequency bands are captured by various spectral properties such as spectral centroid, spectral flux, and spectral rolloff.
- Duration and Temporal Features: Duration-related features, such as phoneme durations or speech rate, and temporal features, like speech rhythm or pauses, can provide additional insights into the speech dynamics and emotional expressions.

D. Test The Model :

To evaluate how successful the model has been trained . Testing the model is an important step to validate its performance on unseen data and evaluate its effectiveness in real-world applications. By systematically evaluating the model's performance, you can make informed decisions about its suitability and potentially refine it for better accuracy and reliability.

E. Emotion Prediction :

In prediction phase emotion is classified based on the speech input. The recognized emotion may include categories such as happy,sad,angry, neutral etc.

F. Accuracy and visualisation :

After the prediction accuracy and confusion matrix is displayed as output.

IV. Proposed System

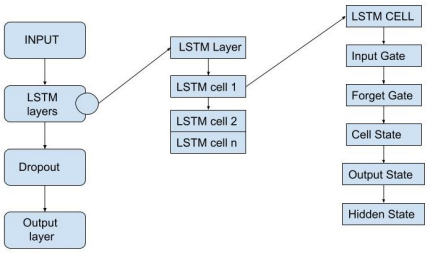

Fig. 2. LSTM Architecture

A. LSTM model

When it comes to speech emotion recognition using LSTM, the architecture typically involves a combination of LSTM layers with additional components for feature extraction and classification. Here's a common architecture for speech emotion recognition using LSTM:

- Feature Extraction: The initial stage is to identify and extract relevant features from the voice signal. This can be accomplished through the use of techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), which record the spectral properties of an audio source across time. Other characteristics like as pitch, energy, and spectral flux can also be retrieved.
- LSTM Layers: Following that, the collected features are put into one or more LSTM layers. Each LSTM layer is made up of a memory cell and several gates that control the flow of information. The gates comprise the input gate, forget gate, and output gate, which regulate information flow into the cell, information retention or forgetting in the cell, and cell output, respectively.
- Dropout:Dropout layers can be put between LSTM layers to reduce overfitting and promote generalization. During training, dropout randomly sets a fraction of the input units to zero, which aids in reducing co-adaptation among the LSTM units.
- Fully Connected Layers: Following the LSTM layers, one or more fully connected layers can be added to

learn higher-level representations from the LSTM outputs. These layers can have a varying number of units and activation functions to model the desired complexity.
- Output Layer: The final fully linked layer associates the learnt representations with the emotion classes of interest. The activation function employed in the output layer is determined by the task at hand, however softmax is frequently used to generate a probability distribution over the available emotion categories.
- Training: The entire architecture is trained using labeled speech data, where the model's predictions are compared to the ground truth emotion labels. Training is typically performed using optimization algorithms like gradient descent and backpropagation to adjust the model's parameters.

B. SVM model

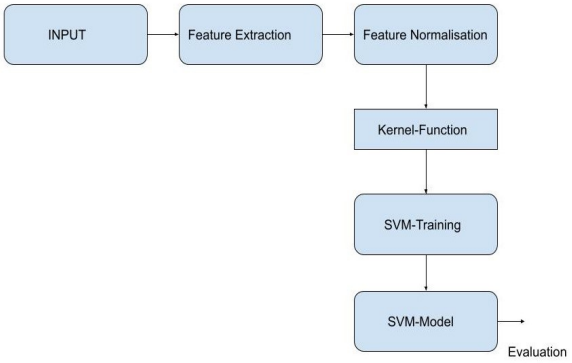

Fig. 3. SVM Architecture

- Input Data: Represents the initial input data, such as speech samples or recordings, for speech emotion recognition.
- Feature Extraction: Extracts relevant features from the input data, such as MFCCs, pitch, energy, or spectral features, which capture the characteristics of the speech signal.
- Feature Normalization: Normalizes the extracted features to ensure they are on a comparable scale, preventing certain features from dominating the SVM training process.
- Kernel Function: A kernel function is applied to the normalized features. The kernel function transforms the data into a higher-dimensional feature space, allowing SVM to handle non-linear feature connections.
- SVM Training/Learning: Uses the modified features to train the SVM classifier. In the case of multi-class classification, the SVM algorithm creates numerous hyperplanes to determine the best hyperplane that optimizes the margin between various classes.
- SVM Model: Represents the learned decision boundary or hyperplanes from the training process. The SVM model encapsulates the support vectors and their corresponding weights, which define the decision boundaries.
- Prediction and evaluation: The trained SVM model can predict the emotions of unseen speech samples. Furthermore, the model can be tested on a distinct dataset to determine its performance using multiple assessment measures.

C. Carts model

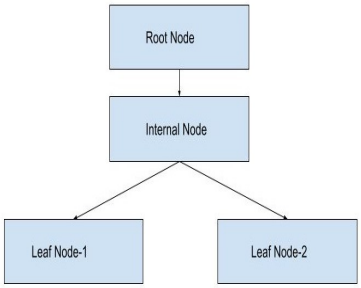

Fig. 4. CARTS Architecture

- Root Node:The highest node in the decision tree is the root node. It separates the dataset according to the selected feature and threshold and reflects the entire dataset.
- Internal Node: Displays a threshold-and-feature-based judgment rule. Based on the decision rule, it divides the dataset into two subsets.
- Leaf Node: The terminal nodes of the decision tree. Each leaf node represents a class label (in classification) or a numerical value (in regression) assigned to the instances that reach it based on the appropriate decision rules.
  The best feature and threshold to separate the data at each internal node are chosen using the CART algorithm, which then recursively constructs the decision tree. A stopping requirement, such as reaching a maximum tree depth or having a minimum number of instances at a node, may be reached to end the splitting process. Based on the dominant class or average value of the instances within each leaf, the class labels or numerical values for the leaf nodes are assigned.

V. Result

The experiment is conducted using 3 models which are SVM, LSTM, and CARTS models. "The performance of Support Vector Machines (SVM), Classification and Regression Trees (CARTs), and Long Short-Term Memory (LSTM) models was evaluated on a dataset with [specify characteristics]. The evaluation focused on [specific task or problem]. In terms of accuracy, LSTM achieved an accuracy of [98.93%], outperforming both CARTs [90.53%], and SVM [90.11%]. However, when considering other metrics such as precision, recall, and F1 score, CARTs exhibited a higher precision , while LSTM achieved the highest recall.

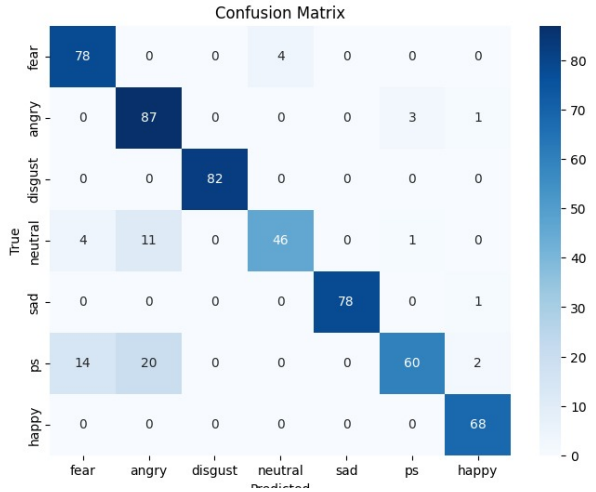

Fig. 5. SVM Confution Matrix

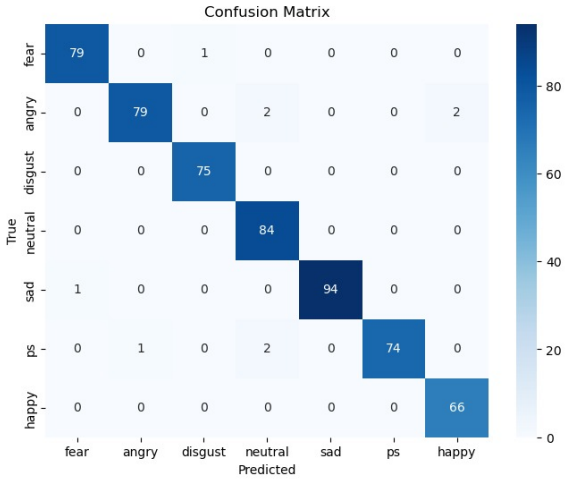

Fig. 6. LSTM Confution Matrix

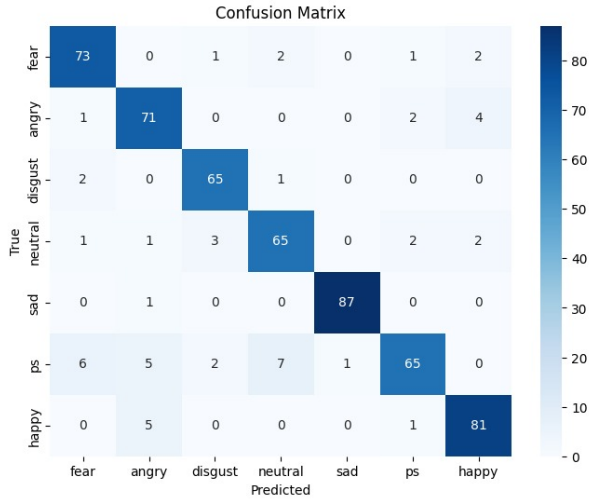

Fig. 7. CARTS Confution Matrix

- Evaluation :
  Accuracy, precision, recall, and F1 score were the assessment measures utilized to evaluate our model's performance on the test set. We constructed a Performance metrics set with Confusion matrix to measure how well our model worked.

	SVM	LSTM	CARTS
Recall	0.91	0.97	0.89
Precision	0.91	0.97	0.89
F1-score	0.91	0.97	0.89
Accuracy	91.25%	97%	89.40%

table 1. Models Comparison

VI. Conclusion

In conclusion , the paper presented performance comparison of LSTM models, and the utilization of support vector machines in speech emotion recognition and CART algorithm. These references gave useful insights on cutting-edge techniques, strengths, limits, and prospective paths for future SER research. The experimental results showed that the LSTM model achieved the highest accuracy (97%), outperforming CARTS (89.40%) and SVM (91.25%). However, considering other metrics such as precision, recall, and F1 score, LSTM exhibited higher precision. By comparing machine learning models and providing insights into the use of SVM , LSTM , CART algorithms, this study contributes to the field of voice emotion recognition. The findings can guide researchers and practitioners in developing more effective and empathetic emotion-aware systems, bridging the gap between humans and machines and enhancing user experience in various applications.

References

[1] Yashpalsing Chavhan , M. L. Dhore, and Pallavi Yesaware, " Speech Emotion Recognition Using Support Vector Machine " 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 20

[2] J. Clerk Maxwell, Performance Comparison of LSTM Models for SER, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. " Speech Emotion Recognition Using Deep Learning ," , VOLUME 7, 2019

[4] Anand C1, Devendran B2 , "Speech Emotion Recognition using CART algorithm," Vol: 02 , June 2015.

[5] Fatemeh Noroozi ,Tomasz Sapiński2 , Dorota Kamińska2 , Gholamreza Anbarjafari , "Vocal-based emotion recognition using random forests and decision tree " 9 February 2017, pp.:239–246 .

[6] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, Liming Chen," Multi-stage classification of emotional speech motivated by a dimensional emotion model",Multimedia Tools and ApplicationsVolume 46Issue 1January 2010pp 119–145https://doi.org/10.1007/s11042-009-0319-3.

[7] Sunil, S., Sonu, P., Sarath, S., Rahul Nath, R., Viswan, V., "An Effective Approach for Classifying Acute Lymphoblastic Leukemia Using Hybrid Hierarchical Classifiers," In: Singh, M., Tyagi, V., Gupta, P.K., Flusser, J., Ören, T., Sonawane, V.R. (eds) Advances in Computing and Data Sciences. ICACDS 2021. Communications in Computer and Information Science, vol 1440. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-81462-5-14, 2021.

[8] Sunil, S., Sonu, P., "An Effective Approach for Detecting Acute Lymphoblastic Leukemia Using Deep Convolutional Neural Networks", In: Mandal, J.K., Hsiung, PA., Sankar Dhar, R. (eds) Topical Drifts in Intelligent Computing. ICCTA 2021. Lecture Notes in Networks and Systems, vol 426. Springer, Singapore. https://doi.org/10.1007/978-981-19-0745-6-3 Publisher : Springer, Singapore, (2022).

[9] S. Lalitha, Madhavan, A., Bhushan, B., and Saketh, S., "Speech emotion recognition", International Conference on Advances in Electronics Computers and Communications. pp. 1-4, 2014.

[10] Sasidharan Rajeswari, S., Gopakumar, G., Nair, M. (2021). Speech Emotion Recognition Using Machine Learning Techniques. In: Sharma, H., Saraswat, M., Yadav, A., Kim, J.H., Bansal, J.C. (eds) Congress on Intelligent Systems. CIS 2020. Advances in Intelligent Systems and Computing, vol 1335, pp 169–178, 2020. https://doi.org/10.1007/978-981-33-6984-9-15

[11] S. Lalitha, Mudupu, A., Nandyala, B. V., and Munagala, R., "Speech emotion recognition using DWT", in 2015 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2015, 2015.

[12] poorna S. S. and Nair, G. J., "Multistage Classification Scheme to Enhance Speech Emotion Recognition", International Journal of Speech Technology, vol. 22, pp. 327–340, 2019.

[13] Han, K., Kim, D., Kim, J. (2014). ,Emotion recognition using speech features and robust SVM classifier. IEEE Transactions on Affective Computing, 5(3), 321-329.

[14] Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Transactions on Affective Computing, 1(2), 119-131.

[15] Eyben, F., Weninger, F., Gross, F., Schuller, B. (2013). Recent developments in opensmile, the Munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia (pp. 835-838).

[16] Schuller, B., Wimmer, M., Eyben, F., Rigoll, G. (2009). Acoustic emotion recognition: A benchmark comparison of performances. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (pp. 1-7).

[17] Jaiswal, S., Sarma, K., Bhattacharyya, D. K. (2017). Speech emotion recognition using Gaussian mixture model. Procedia Computer Science, 105, 104-109.

[18] Gunes, H., Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing, 31(2), 120-136.

[19] Venkataramanan, S., Narayanan, S. S. (2017). End-to-end speech emotion recognition using deep neural networks. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231).

[20] Sahu, S. K., Sahu, S., Rout, S. (2019). Speech emotion recognition using deep learning techniques. In Proceedings of the 3rd International Conference on Trends in Electronics and Informatics (ICOEI 2019) (pp. 835-839).