A Comparative Analysis of Machine Learning and Deep Learning Techniques for Speech Emotion Recognition

VINAL WAGHELA[1], DR. MIRAL PATEL[2], PROF. RAHUL PATEL[3]
[1, 2, 3] *G.H Patel college of engineering*

Abstract*—This review paper explores Speech Emotion Recognition (SER) techniques, comparing traditional Machine Learning (ML) and Deep Learning (DL) approaches. We analyze the strengths and weaknesses of each approach, particularly regarding feature engineering and model complexity. The paper discusses how ML methods rely on handcrafted features like Mel-Frequency Cepstral Coefficients (MFCCs) for emotion classification using algorithms like Support Vector Machines (SVMs) or k-Nearest Neighbors (kNN). Conversely, Deep Learning techniques, particularly Long Short-Term Memory (LSTM) networks, have emerged as powerful tools for SER due to their ability to automatically learn features directly from raw audio data. We examine the trade-offs between interpretability of ML models and the data-driven feature learning capabilities of Deep Learning. Additionally, the paper explores challenges faced by both approaches, including data availability and domain adaptation. Finally, we discuss the potential applications of SER technology across various domains and highlight promising future directions in this evolving field.

***Index Terms*— Long Short-Term Memory (LSTM), Mel-frequency cepstral coefficients (MFCCs), Toronto Emotional Speech Synthesis (TESS) dataset**

I. INTRODUCTION

Imagine a world where machines can understand not just the words we say, but also the emotions we convey. This is the promise of speech emotion recognition (SER), a field in artificial intelligence (AI) that aims to crack the code of human emotions hidden within spoken language. While traditional methods relied on hand-crafted features and algorithms, the rise of deep learning has brought a revolution. Deep learning architectures, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are masters at uncovering hidden patterns in vast amounts of data. This makes them perfect for SER, where emotions are often subtly woven into the intricate fabric of speech. Virtually all the ASR algorithms and services are simply transcribing audio recordings into written words. But that is only the first level of speech understanding. During the conversation humans receive lots of meta-information apart from text. Examples might be the person who is speaking, his intonation and emotion, loudness, shades etc. These factors might considerably influence the true intended meaning of a phrase. Even turn it into opposite - that is what we call sarcasm or irony. Humans take all these elements into consideration while processing the phrase in the brain and only after that the final meaning is formed.

Nowadays machines can successfully recognize human speech. Automatic speech recognition (ASR)services can be found everywhere. Voice input interfaces are used for many applications from navigation system in mobile phones to Internet-of-Things devices [1]. As one of the most fundamental characteristics that distinguishes intelligent life forms from the rest, emotion is an integral part of our daily conversations. From the broad perspective of general-purposed artificial intelligence, the ability to detect the emotional contents of human speech has far reaching applications and benefits. Furthermore, the notion that machines can understand and perhaps some day produce emotions can profoundly change the way humans and machines interact [2]. An increasing number of people will interact with a voice assistance machine than with their partners in the next five years. With proliferation of Virtual Personal Assistants (VPA) such as Siri, Alexa and Google Assistant in our day-to-day interactions, they fill a role of answering our questions and fulfilling our requests quickly and accurately. Though these assistants understand our commands, they are not proficient enough in recognizing our mood and reacting accordingly. Therefore, it is pertinent to develop an efficient emotion recognition system which can enhance the capabilities of these assistants and revolutionize the whole industry [3]. Speech is a rich, dense form of communication that can convey information

effectively. It contains two types of information, namely linguistic and paralinguistic. The former refers to the verbal content, the underlying language code, while the latter refers to the implicit information such as body language, gestures, facial expressions, tone, pitch, emotion etc. Para linguistic characteristics can help understand the mental state of the person (emotion), gender, attitude, dialect, and more [4].Recorded speech has key features that can be leveraged to extract information, such as emotion, in a structured way. To get such information would be invaluable in facilitating more natural conversations between the virtual assistant and the user since emotion color everyday human interactions [2]. This review paper conducts a comparative analysis of Machine Learning (ML) and Deep Learning (DL) techniques for Speech Emotion Recognition (SER).

II. TRADITIONAL MACHINE LEARNING TECHNIQUES FOR SER

In traditional Machine Learning (ML) for Speech Emotion Recognition (SER), feature engineering plays a critical role. It's the process of transforming raw audio data into a meaningful representation that machine learning algorithms can understand and use for emotion classification. Here's a breakdown of why feature engineering is important and how it works:

Importance of Feature engineering: Speech signals are complex waveforms containing information about pitch, volume, and spectral content (how sound is distributed across frequencies). Directly feeding this data into an ML model wouldn't be effective. Feature engineering extracts relevant characteristics that relate to emotions. Speech data contains a vast amount of information, not all of which is relevant for recognizing emotions. Feature engineering allows you to highlight aspects of the speech signal that are likely to carry emotional cues. This helps the model focus on the important parts of the data for accurate classification. Traditional ML algorithms require structured, numerical data as input. Feature engineering acts as a bridge, transforming the raw audio data into numerical features (e.g., MFCCs, pitch) that the algorithms can use for learning and classification.

Feature Engineering Process :

Figure 1 Speech Signal Analysis *[5]*

Data Preprocessing, This involves cleaning the audio data by removing noise, silence, or irrelevant portions. Feature Extraction this step involves applying various techniques to extract specific characteristics from the pre-processed audio. An important issue in the design of a speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions. Since pattern recognition techniques are rarely independent of the problem domain, it is believed that a proper selection of features significantly affects the classification performance. An important issue in speech emotion recognition is the extraction of speech features that efficiently characterize the emotional content of speech and at the same time do not depend on the speaker or the lexical content. Although many speech features have been explored in speech emotion recognition, researchers have not identified the best speech features for this task [6]. Common features used in SER include Mel-Frequency Cepstral Coefficients (MFCCs), Pitch (Fundamental Frequency), Prosodic Features.

Mel-frequency cepstrum coefficient (MFCC) is the most used representation of the spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the Mel-frequency scale [7]. MFCCs represent the spectral shape of the audio signal on a mel scale, which approximates human auditory perception. They capture characteristics like formants (resonant frequencies of the vocal tract).

Prosodic Feature: The prosodic features are categorized into 5 types as mentioned below: I. Intensity or Energy contour features 2. Pitch or Fundamental frequency contour features 3. Prosodic Spectral features 4. Formant frequency contour features and 5. Additional acoustic features. [8]

Encompass characteristics related to speech rhythm and intonation, such as energy (intensity), zero-crossing rate (frequency of crossing zero amplitude), and speech rate. These can reflect emotional variations like increased energy in anger or slower speech rate in sadness.

Finally the last step of feature engineering process is classification. Majority of the studies applied standard classifiers such as SVM (Support Vector machine) , GMM (Gaussian Mixture model) , K-NN (K Nearest Neighbor), HMM (Hidden Markov Model).

Support Vector Machines (SVMs):
Powerful classifiers that learn decision boundaries to separate different emotions based on extracted features like MFCCs (spectral features) and pitch. SVM is a very simple and efficient classifying algorithm which is used for classification and pattern recognition. Support Vector Machines algorithm was introduced by Vladimir Vapnik in 1995.The main aim of this algorithm is to obtain a function that constructs hyper planes or boundaries. These hyper planes are used to separate different categories of input data points. There are typically two types of SVM classifiers, Linear and Nonlinear. Also, SVM has kernel functions that can be used in supervising environment. In training phase, we are using radial basis kernel function because it limits the training datasets to lie within the specified boundaries. LIBSVM tool is used for SVM classification. During training of signal, feature values of speech signals that are extracted from speech signals are send to LIBSVM with their class labels as Happy, Angry, Sad, and Fear. SVM model is obtained for each emotional state by using their feature values which are extracted. Once the training model has been prepared with, it is easy to predict the emotional states with testing datasets. Features are extracted from speech signals and with the help of SVM model values generated by training models, the emotions are classified automatically as Happy, Angry, Sad or Fear [9].

Gaussian Mixtures models:
A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model. A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$P(x/\lambda) = wi \; M \; i=1 \; g(x \; \mu i \; , \; i \;)$ …… (1)

Where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), wi, i = 1, . . . ,M, are the mixture weights, and

$g(x \; \mu i \; , i \;) \; , i = 1, \ldots \; ,$

M are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$g(x \; \mu i \; , i \;) = 1 \; 2\pi \; D/2 \; | \; i \; | \; 1 \; 2 \; \exp\{- 1 \; 2 \; x - \mu i \; ' \; x - \mu i \; -1 \; i \; \}$ …… (2)

With mean vector μi and covariance matrix . i the mixture weights satisfy the constraint that

$wi \; M \; i=1 = 1$ the complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation [10].

K-Nearest Neighbors (KNN):

Classifies new data points based on the majority vote of its k nearest neighbors in the feature space, where neighbors are determined by feature similarity. In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the future space. The output depends on whether k-NN is used for classification or regression. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. A commonly used distance metric for continuous variables is Euclidean distance. In our research, we use the system to extract the speech's feature. After the feature extraction, we give each speech sample with the corresponding emotion class label. After that we input them to the LIB SVM

classifier and gain a model file by training the data set. When an unclassified speech sample come into this system, the system extract the feature coefficients and use the model file to classify the speech emotion [11].

Hidden Markov Models (HMM):
Model speech signals as sequences of hidden states, each representing a potential emotional state. HMMs use probabilities to transition between states and emit specific features. Hidden Markov models (HMMs) are popular for speech recognition (Lee and Hon, 1989) and hence they are adopted for the classification of emotion in speech. According to Deller et al. (1993), the states in the HMM frequently represent identifiable acoustic phonemes in speech recognition. The number of states is often chosen to roughly correspond to the expected number of phonemes in the utterances. However, the optimal number of states is best determined through experiments as the relationship of the number of states to the performance of the HMM is very imprecise [12].

Challenges of Feature Engineering:
An important issue in the design of a speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions. Since pattern recognition techniques are rarely independent of the problem domain, it is believed that a proper selection of features significantly affects the classification performance. Four issues must be considered in feature extraction. The first issue is the region of analysis used for feature extraction. While some researchers follow the ordinary framework of dividing the speech signal into small intervals, called frames, from each which a local feature vector is extracted, other researchers prefer to extract global statics from the whole speech utterance. Another important question is what the best feature types for this task are, e.g. pitch, energy, zero crossing, etc.? A third question is what is the effect of ordinary speech processing such as post-filtering and silence removal on the overall performance of the classifier? Finally, whether it suffices to use acoustic features for modelling emotions or if it is necessary to combine them with other types of features such as linguistic, discourse information, or facial features [6].

1. Domain Knowledge Required:
Choosing the most effective features for your ML model depends heavily on the specific task and dataset you're working with. In SER (Speech Emotion Recognition), for instance, selecting the right features requires expertise in both speech processing and the domain of emotion recognition. This can be a barrier for researchers or practitioners who lack experience in these specific areas.

2. Time-Consuming and Iterative Process:
Feature engineering is often an exploratory process that involves experimentation with different feature extraction techniques and selection methods. It requires iteratively evaluating the impact of different feature sets on the performance of your ML model. This can be time-consuming, especially for complex tasks with a large number of potential features.

3. Limited Generalizability:
Features engineered for one specific dataset or task might not be directly transferable to another domain with different characteristics. For example, features effective for recognizing emotions in acted speech data from professional actors might not be optimal for recognizing emotions in spontaneous speech recordings. This necessitates re-engineering features when applying the model to a new dataset or domain.

4. Feature Selection Complexity:
Extracting a large number of features might not always be beneficial. Irrelevant or redundant features can increase the computational cost of training the ML model and potentially lead to overfitting. Feature selection techniques can help address this by identifying the most informative features that contribute the most to the model's performance. However, choosing the right feature selection method itself can be a challenge, requiring careful consideration of the specific task and dataset.

5. Interpretability:
Traditional ML models with handcrafted features can be easier to interpret compared to complex deep learning models. Understanding the features used by the model can provide insights into the decision-making process and identify potential biases. Feature engineering allows for some level of control over the

interpretability of the model by selecting features that are more readily interpretable.

III. DEEP LEARNING TECHNIQUES FOR SER

Deep Learning has been considered as an emerging research field in machine learning and has gained more attention in recent years [13] [14]. Deep Learning techniques for SER have several advantages over traditional methods, including their capability to detect the complex structure and features without the need for manual feature extraction and tuning; tendency toward extraction of low-level features from the given raw data, and ability to deal with un-labelled data [14]. Unlike traditional Machine Learning (ML) that relies on handcrafted features for tasks like SER (Speech Emotion Recognition), deep learning models automatically learn features directly from the raw audio data. Deep learning models use artificial neural networks with multiple interconnected layers. Each layer performs a specific transformation on the data, ultimately aiming to extract meaningful representations. The raw audio data, a complex waveform representing speech, is fed as input to the first layer of the network.

Deep learning methods are comprised of various nonlinear components that perform computation on a parallel basis [15] [14]. However, these methods need to be structured with deeper layers of architecture to overcome the limitations of other techniques. Deep learning techniques such as Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Networks (CNN) and Auto Encoder (AE) are considered a few of the fundamental deep learning techniques used for SER, that significantly improves the overall performance of the designed system. Deep learning is an emerging research field in machine learning and has gained much attention in recent years. A few researchers have used DNNs to trained their respective models for SER [14]. In this section, we provide an overview of the various deep learning techniques utilized for SER.

CNN (Convolutional Neural Network): The architecture of CNN is shown in Fig. 2, which has an input layer, one convolutional layer, one fully connected layer, and a SVM classifier. We use the spectrogram of the speech signal as the input of CNN. The main idea of feature learning is to learn high-level representations from the low-level raw features, and the spectrogram is well-suited for this task. As a low-level feature, spectrogram is widely used in speech recognition and audio-based speaker and gender recognition. Following the hierarchy of CNN, the features learned at each layer become increasingly invariant to nuisance factors while maintaining affect-salience with respect to the goal of SER [16].

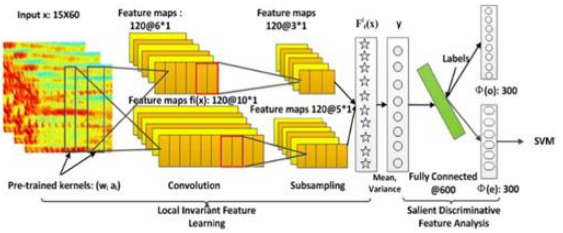

Figure 2. System pipeline. Left: Input spectrogram at two different resolutions. The next stage is the local invariant feature learning containing the output of one long feature vector. The salient discriminative feature learning produces the last stage of affect-salient features and nuisance features then fed to a linear SVM for SER. *[16]*

Convolutional Neural Networks (CNNs) have become a powerful tool for Speech Emotion Recognition (SER) due to their ability to extract informative features directly from raw audio data. They are often used as building blocks in more complex deep learning architectures that combine CNNs with RNNs to leverage both spatial and temporal information for improved emotion recognition.

Recurrent Neural Network:
Recurrent Neural Networks (RNNs) have emerged as a powerful tool for Speech Emotion Recognition (SER) due to their ability to capture the sequential nature of speech data. Unlike humans who can perceive emotions throughout an utterance, traditional machine learning methods often struggle with this temporal aspect.

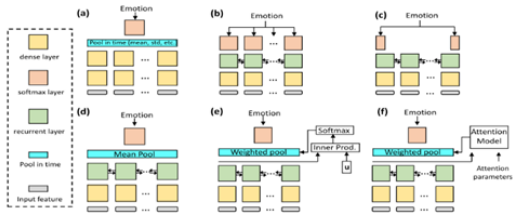

Figure 3 Architectures for applying DNN/RNN for SER *[17]*

Most of the features listed in Figure 3 can be inferred from a raw spectrogram representation of the speech signal. It is therefore reasonable to assume that given a fixed set of (differentiable) HSFs and sufficient data, similar short-term features can be learned from a raw spectral representation. Figure 3 shows an example structure to learn short-term LLDs, using a few layers of dense nonlinear transformations. Note that the statistical functions in the context of neural networks function as pooling layers over the time dimension [17]. Why RNNs Excel in SER is because of Modelling Temporal Dependencies Speech is not a static snapshot, but rather unfolds over time with variations in pitch, intensity, and speaking style. RNNs excel at capturing these temporal relationships within sequences, making them ideal for understanding how emotional cues evolve throughout an utterance. Handling Variable Length Input: RNNs can effectively process speech segments of varying lengths, unlike some methods requiring fixed-size inputs. This flexibility is crucial for real-world scenarios where speech patterns can differ significantly.

Long Short term memory (LSTM) : Long Short-Term Memory (LSTM) networks, a specific type of Recurrent Neural Network (RNN), have become a dominant force in Speech Emotion Recognition (SER) due to their exceptional ability to capture the temporal dynamics of emotions within speech. The LSTM model [18] [19]. is a powerful recurrent neural system specially designed to overcome the exploding/vanishing gradient problems that typically arise when learning long-term dependencies, even when the minimal time lags are very long [18] [19].Overall, this can be prevented by using a constant error carousel (CEC), which maintains the error signal within each unit's cell. As a matter of fact, such cells are recurrent networks themselves, with an interesting architecture in the way that the CEC is extended with additional features, namely the input gate and output gate, forming the memory cell [19].

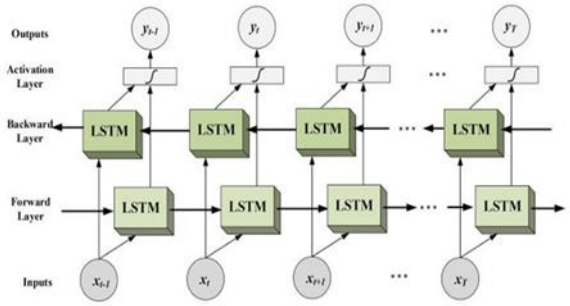

Figure 4 Using LSTM model

Attention mechanism: Attention mechanisms have emerged as a powerful tool for improving Speech Emotion Recognition (SER) by allowing deep learning models to focus on the most relevant parts of the speech input for emotion recognition. The Attention Mechanism is widely used to improve the performance of SER [20]. Speech is a complex signal, and emotions can be expressed through subtle variations in pitch, intensity, and speaking style throughout an utterance. Traditional deep learning models like CNNs and RNNs process the entire speech sequence equally. However, not all parts of the speech hold the same weight in conveying emotion. Attention mechanisms address this challenge by introducing a dynamic weighting scheme. Here's how it works:

Learning Feature Representations: The model first processes the speech input using techniques like CNNs or LSTMs, generating a sequence of feature vectors representing different segments of the speech.
Attention Scores Calculation: An attention layer takes these feature vectors as input and computes an "attention score" for each one. This score reflects how relevant each segment is for recognizing the overall emotion. Factors like pitch variations during emphasis or changes in speaking style during excitement might influence the attention scores.

Weighted Representation: Based on the attention scores, the model creates a weighted representation of the entire speech sequence. Segments with higher attention scores contribute more to the final representation, effectively focusing on the most

informative parts of the speech for emotion recognition.

Emotion Prediction: Finally, the model uses this weighted representation to predict the most likely emotion for the speech segment.

IV. COMPARING MACHINE LEARNING AND DEEP LEARNING FOR SER

Table 1 . difference between deep learning and machine learning for SER

Sr no.	Evaluation Criteria	Deep learning	Machine learning
1.	Feature extraction	Automatic feature extraction directly from raw audio data (e.g., MFCCs) using techniques like CNNs.	Requires manual feature engineering, which can be time-consuming and domain-specific.
2.	Performa-nce	Generally achieves higher accuracy in emotion recognition by capturing complex patterns in speech data.	Performance can be limited by the quality of handcrafted features.
3.	Adaptability	Can potentially adapt to new data or variations in speech due to continuous learning during training.	May require feature re-engineering for adapting to new data or speech characteristics.
4	Interpretability	Can be a "black box," making it difficult to understand the reasoning behind emotion predictions	Some models offer better interpretability, allowing insights into feature importance for emotion recognition.

The above given Table 1 highlights the key differences between deep learning and machine learning for SER, showcasing deep learning's advantages in automatic feature extraction, performance, and adaptability, while acknowledging machine learning's potential for interpretability.

Name: Speech Emotion Recognition Using Support Vector Machine [9] [21].
Technology used: Basic SVM
Strong point: MFCC+MEDC+Energy features give the best accuracy among various spectral and prosodic features [21].
Weak point: If this system applied to actual-time SER then accuracy is not satisfactory [21].

Name : Efficient Speech Emotion Recognition using Binary support Vector Machines & Multiclass SVM [22] [21].
Technology used: Binary SVMs and Multi-Class SVMs.
Strong point: Higher accuracy of MFCC features using Linear and RBF kernels with 3-stage hierarchical SVM [21]
Weak point: Only MFCC features are considered as well as the sigma value is also set to lowes [21].

Name: Speech Emotion Recognition using Convolutional and Recurrent neural networks [23] [21]
Technology used: Time Distributed CNN , RNN
Strong point: It applies the same layer to several inputs. So, it is efficient and saves time [21].
Weak point: It provides accuracy though when combined with other approach it would be more useful [21].

Name: Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network [23] [21].
Technology used : Deep CNN-LSTM.
Strong point: Short and discriminative features are learning automatically which is a basic step for SER [21].
Weak point: In case of fear and happy emotions it often gets confused among the spectrograms [21].

The above comparison is of various approaches for Speech Emotion Recognition (SER), showcasing the evolution from traditional machine learning techniques using handcrafted features (e.g., SVM with MFCC features) to deep learning models (e.g., Deep CNN-LSTM) that automatically learn informative features from raw speech data. While machine learning approaches can achieve good accuracy, deep learning offers the advantage of automatic feature extraction and potentially higher performance, though interpretability can be a challenge.

In Speech Emotion Recognition (SER), machine learning and deep learning approaches offer distinct advantages and disadvantages. Machine learning methods, like Support Vector Machines (SVMs), often rely on handcrafted features (e.g., MFCCs) which can be time-consuming to engineer and may not capture the full complexity of emotions in speech. While interpretability can be a benefit of some machine learning models, their performance can be limited by the quality of these features. Deep learning techniques, on the other hand, automatically learn features directly from raw audio data using architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This eliminates the need for manual feature engineering and allows deep learning models to capture intricate patterns in speech, potentially leading to higher recognition accuracy. However, deep learning models can be computationally expensive to train and may be less interpretable compared to their machine learning counterparts. Choosing the best approach depends on the specific needs of the application, with deep learning generally favored for superior performance when computational resources and interpretability are not paramount concerns.

V. APPLICATION OF SPEECH EMOTION RECOGNITION

[24]proposed a real-time emotional state recognition system for mobile phones. The system utilizes an INTEL Dialogic D/4PCI board for sound capture and classifies emotions into five categories: neutral, happy, sad, angry, and annoyed. Each classification comes with an associated probability. To mitigate mobile noise and enhance performance, the authors implemented a Moving Average (MA) filter. Furthermore, they employed Sequential Forward Selection (SFS) for feature optimization and stability. For classification, they opted for a Support Vector Machine (SVM) with probability estimation capabilities [8] and compared its performance to a k-Nearest Neighbors (k-NN) algorithm. Finally, the authors envisioned a practical application: a mobile agent that gauges affection levels (love, truthfulness, weariness, trickery, friendship) during phone conversations. This agent highlights the potential of the system for various mobile applications.

[25]Their work investigates the role of emotion recognition in enhancing e-learning systems. The authors emphasize the importance of considering a learner's emotional state during the learning process. They explore various methods for emotion recognition, including self-reporting, physiological data analysis, and facial expression detection. The research concludes that a multimodal approach, combining multiple recognition techniques, yields the most effective results for e-learning applications.

study done by [26] explores using speech patterns to detect depression. They analyzed speech from depressed and healthy people in different situations (interviews, describing pictures, reading) and with different emotions (positive, neutral, negative). Their method achieved promising accuracy (around 80%) in identifying depression from speech patterns.

Deep learning has significantly advanced the field of Speech Emotion Recognition (SER), opening doors to its application in a wide range of domains. In Human-Computer Interaction (HCI), SER can improve user experience by enabling systems to adapt to a user's emotional state. Customer service can benefit from SER by allowing for better identification of customer sentiment and tailoring interactions accordingly. Education and learning applications can leverage SER to gauge student engagement and potentially personalize instruction. Mental health monitoring and care can potentially utilize SER to detect signs of depression or anxiety through speech patterns. Security and public safety applications could benefit from SER by identifying emotional cues that might indicate threats or distress. The field of entertainment and gaming can leverage SER to create more immersive and emotionally responsive experiences.

Market research and product development can potentially utilize SER to gain insights into customer sentiment towards products or services. Finally, robotics and social assistive technology can benefit from SER by enabling robots to better understand and respond to human emotions. This diverse range of applications highlights the transformative potential of deep learning-powered SER.

CONCLUSION

This review paper has explored the landscape of Machine Learning (ML) and Deep Learning (DL) techniques for Speech Emotion Recognition (SER). We have examined traditional ML algorithms like Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN) alongside Deep Learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The comparison highlighted the strengths and weaknesses of each approach. While ML methods offer interpretability and efficiency, they often struggle with complex feature engineering and limited data handling capabilities. Deep Learning, on the other hand, excels at feature extraction and achieves higher accuracy with large datasets. However, DL models can be computationally expensive and lack interpretability. This conclusion emphasizes Strengths and weaknesses of ML and DL for SER ,Importance of interpretability and data handling, Potential for, hybrid approaches and explainable AI, Diverse applications of SER technology, Importance of ethical considerations

REFERENCES

[1] P. P. Vladimir Chernykh, "Emotion Recognition From Speech With Recurrent Neural Network," Arxiv, vol. 2, p. 18, 5th July 2018.

[2] M. X. ,. R. C. ,. E. D. ,. J. D. ,. V. T. Jianyou Wang, "Speech Emotion Recognition With Dual-Sequence LSTM Architecture," ICASSP 2020, p. 5, 2020.

[3] H. R. R. Kannan Venkataramanan, "Emotion Recognition from Speech," arxiv, vol. 1, p. 14, 22 december 2019.

[4] Y. Yamashita, "A review of paralinguistic information processing for natural speech communication," Acoustical Science and Technology, p. 7, 2013.

[5] ∗. ,. T. C. a. ,. O. A. a. ,. J. M. T. b. ,. C. P. c. ,. D. P. d. ,. S. L. S. Samaneh Madanian a, "Speech emotion recognition using machine learning — A systematic review," Elsevier, p. 25, 2023.

[6] ,. M. S. K. b. Moataz El Ayadi a, "Survey on speech emotion recognition: Features, classification schemes, and databases," Elsevier, p. 16, March 2011.

[7] Y. S. M. M. K. R. M. A. M. e. a. Leila Kerkeni, "Automatic Speech Emotion Recognition Using Machine Learning," HAL Open science, p. 18, 2019.

[8] A. R. K. N. K. Monorama Swain, Study of prosodic feature extraction for multidialectal Odia Speech emotion recognition, 2016 IEEE Region 10 Conference (TENCON), 2016.

[9] S. N. P. B. B. K. P. A. B. K. R. K. M. Manas Jain, "Speech Emotion Recognition using Support Vector Machine," arxiv, 3 february 2020.

[10] P. C. A. P. M. D. D. Patel, Proceedings of the International Conference on Science & Engineering for Sustainable, p. 9, May 2017.

[11] G. B. K. M. S. G. G. Anuja Bombatkar, "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm," International Journal of Engineering Research and Applications, 20144.

[12] S. W. F. L. C. D. S. Tin Lay Nwe, "Speech emotion recognition using hidden Markov models,," Elsevier, vol. 41, no. 1, 30 June 2003.

[13] J. Schmidhuber, "Deep learning in neural networks: An overview,," Elsevier, vol. 61, 2015.

[14] E. J. M. I. B. T. J. M. H. Z. a. T. A. R. A. Khalil, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," vol. 7, 2019.

[15] A. B. S. S. D. S. Björn Schuller, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,," Elsevier, p. 27, 2011.

[16] M. D. Z. H. a. Y. Z. Q. Mao, "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," IEEE

Transaction on multimedia, vol. 16, p. 11, december 2015.

[17] E. B. a. C. Z. S. Mirsamadi, "Automatic speech emotion recognition using recurrent neural networks with local attention,," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 5, 2017.

[18] S. H. a. J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, pp. 1735-1780, 1997.

[19] G. M. C. &. N. Van Houdt, "A review on the long short-term memory model," Artificial Intelligence Review, p. 27, 2020.

[20] Z. M. Y. X. Z. Z. Z. T. S. X. Chen S, "The Impact of Attention Mechanisms on Speech Emotion Recognition," p. 20, 12 November 2021.

[21] S. P. P. S. a. M. R. A. Sandesara, "A Comparative Study On Speech Emotion Recognition," IJRESM, vol. 3, p. 11.

[22] N. R. K. a. S. Saraswathi, "Efficient speech emotion recognition using binary support vector machines & multiclass SVM," 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), p. 6, 2015.

[23] J. A. N. R. a. S. W. B. A. M. Badshah, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 International Conference on Platform Technology and Service (PlatCon), p. 5, 2017.

[24] W. C. Y. P. K. Yoon, "A Study of Speech Emotion Recognition and Its Application to Mobile Services," Indulska, J., Ma, J., Yang, L.T., Ungerer, T., Cao, J. (eds) Ubiquitous Intelligence and Computing., p. 10, 2017.

[25] G. A. M. Maryam Imani, "A survey of emotion recognition methods with emphasis on E-Learning environments,," Journal of Network and Computer Applications, 2019.

[26] B. H. Z. L. L. Y. T. W. F. L. H. K. X. L. Haihua Jiang, "Investigation of different speech types and emotions for detecting depression using different classifiers,," Speech Communication, p. 8, 2017.