Advanced Techniques for Recognizing Emotions: A Unified Approach using Facial Patterns, Speech Attributes, and Multimedia Descriptors

Kummari Ramyasree[1], Chennupati Sumanth Kumar[2]
Department of E & ECE, GITAM Deemed to be University, Visakhapatnam -530045, AP, India[1, 2]
Department of ECE, Guru Nanak Institutions Technical Campus, Hyderabad -501506, Telangana, India[1]

Abstract*—The inability to efficiently store distinguishing edges, local appearance-based textured descriptions generally have limited performance in detecting facial expression analysis. The existing technology has certain drawbacks, such as the potential for poor edge-related disturbance in face photos and the reliance on present sets of characteristics that might fail to adequately represent the subtleties of emotions and thoughts in a variety of contexts. In order to overcome the difficulties associated with identifying facial expressions identification and emotion categorization, this study presents an innovative structure that combines three different information sets: a new multimedia descriptors, prosodic functions, and Local Differential Pattern (LDP). The principal driving force is the existence of noise-induced warped and weak edges in face pictures, which result in inaccurate expressions characteristic assessment. By identifying and encoding only greater edge reactions, as opposed to standard local descriptors that the LDP approach improves the endurance of face feature extraction. Robinson Compass and Kirsch Compass Masks are used for recognising edges, and the LDP formulation encodes each pixel with seven bits of information to reduce code repetition. The last category comprises Long-Term Average Spectrum (LTAS) obtained from signals related to speech, Mel-Frequency Cepstral Coefficients (MFCC), and Forming agents. Fisher Criterion is used to reduce dimensionality, and unpredictable characteristics are used in picking features. Emotion prediction is achieved by classifying two distinct circumstances using Support Vector Machine (SVM) and Decision Tree (DT) algorithms, and combining the obtained data. The research also presents a unique Visual or audio Descriptors that gives priority to key structure selections and face positioning for Audio-visual input. A concise depiction of expression is offered by the suggested Self-Similarity Distance Matrix (SSDM), which uses facial highlight points to estimate both time and space correlations. Formant frequency range, energy sources, probabilistic properties, and spectroscopic aspects define the acoustic signal. The 98% accuracy rate is attained by the emotion recognition algorithm. Major improvements upon cutting-edge techniques are shown in validation studies on the SAVEE and RML information sets, highlighting the usefulness of the suggested model in identifying and categorising emotions and facial movements in a variety of contexts. The implementation of this research is done by using Python tool.

Keywords*—*Local Difference Pattern (LDP); Mel-Frequency Cepstral Coefficients (MFCC); Long-Term Average Spectrum (LTAS); Self-Similarity Distance Matrix (SSDM); Support Vector Machine (SVM)

I. INTRODUCTION

Emotion recognition is an intricate yet essential component of human communication, and computing and AI has placed a great deal of emphasis on it recently. The capacity to precisely recognise and comprehend human feelings has broad applications, from boosting behavioural care providers to advancing consumer-computer connection [1]. Facial expressions are an inherent means for people to communicate a wide range of emotions, and algorithmic systems may be developed to read similar nuanced signs [2]. Basic features including mouth expressions, blinking eyes and eyebrows twitches may be detected by facial identification applications, which can then be used to deduce moods like surprise, pleasure, sorrow, and rage. Apart from expressions on the face, speech assessment is an essential component in the identification of emotions [3][4]. Another approach used in mood detection systems is identification of gestures. Human behaviour, which includes posture and movements, may give important information about how one feels. Huge collections of movement of the body training may be used to train machine learning algorithms to recognise certain gestures as indicative of certain feelings [5]. This method is especially useful in situations when it may be difficult to read or understand facial emotions, including in crowded spaces or through video conferences. The creation of emotions detection mechanisms has great potential for a variety of uses as technology evolves. The capacity of computers to comprehend and react to human emotions offers up fresh prospects for developing empathic and flexible innovations, from interactions between humans and computers to mental wellness assessment and personalised experiences for consumers [6]. Still, the significance of appropriate research and implementation in such developing sector is underscored by ethical issues, privacy problems, and the possibility of bias in identifying feelings systems.

Numerous methodologies are now in use for recognising feelings, and they all use different forms and ways to gather and process emotional information. Among the techniques that have been studied and used frequently is facial expressions assessment. This method entails the detection and analysis of facial reactions and motions employing machine learning tools. Convolutional Neural Networks (CNNs) are a popular neural network method used by investigators and programmers to recognise similarities linked to different feelings from enormous datasets of labelled facial movements

[7]. By examining the meaning included in words written or spoken, machine learning techniques improve this technique even further and make it possible to identify the feelings communicated via language. Voice-enabled devices, digital assistants, and contact centres are among the places where identifying emotions in conversation is frequently used [8]. The field of gesture identification is centred on the interpretation of nonverbal cues such as posture, movements of the hands, and various other physiological motions. This technique is especially useful in situations when facial emotions might not be apparent or may lack sufficient data. Machine learning techniques may be developed on databases that link particular movements to associated sentiments [9]. Such algorithms frequently depend on neural networks with recurrence or related sequence-based designs. Applications such as augmented realities, surveillance footage, and interactions between humans and computers can benefit from this strategy. The use of multiple methods techniques work especially well in everyday life when there are a variety of influenced by context signs of emotion [10]. Even though these techniques have shown promise in a number of applications, problems still exist. Important factors to take into account are the possibility of biased results, confidentiality difficulties, and ethical issues.

A significant barrier to attaining global application is the subtleties and cultural differences in expressing one's feelings [11]. Dependent on environment is a further important restriction. Because feelings are so dependent on the setting, current frameworks may find it difficult to account for the subtle differences in how emotions seem in various contexts. The subjective nature and variability among individuals present further difficulties for emotion detection methods trying to be generalised [12]. Individualised variations in personality, socioeconomic situation, or mental state can provide unpredictability that is challenging for systems to compensate for, and humans may exhibit moods in various manners. The use of such devices in vulnerable or public areas may violate people's fundamental right to safety, hence rules and moral requirements must be carefully considered in order guarantee proper usage [13].

Recent years have seen tremendous progress in the field of emotion recognition, especially with the incorporation of multimedia descriptors, speech characteristics, and facial patterns. Despite these advancements, there are still significant obstacles in the way of the practical uses of current techniques. A significant problem is the restricted capacity to represent the subtleties of affective responses in many media. The process of feature extraction, which is essential for distinguishing minute differences in speech intonations, facial expressions, and multimedia information, is frequently difficult for traditional approaches. The suggested strategy addresses the drawbacks of current techniques by introducing a number of novel algorithms in response to these difficulties. We highlight how important it is to extract features using the Local Directional Pattern (LDP) method, which improves facial pattern representation and guarantees a deeper examination of emotional indicators. But the problem goes beyond feature extraction, which is a precisely sophisticated method like feature selection and dimensionality reduction are needed. In order to ensure a more effective and efficient emotion detection procedure, these stages are essential for simplifying the data and conserving just the most discriminative aspects.

The limitations of traditional classification techniques are addressed by our approach's use of Decision Tree Classifier and Support Vector Machines (SVM). These classifiers improve the system's capacity to identify intricate patterns in multimedia information, speech tones, and facial expressions, enabling a more complicated comprehension of emotions. Moreover, the incorporation of the Self-Similarity Distance Matrix (SSDM) enhances the analysis and facilitates a thorough assessment of similarity patterns among modalities. A fundamental problem with current emotion identification systems is the absence of a cohesive methodology that effectively integrates data from many sources. Seeing this gap, the foundation of our strategy is our suggested Emotion Recognition Fusion mechanism. This fusion technique offers a comprehensive and more accurate representation of the subject's emotional state by cleverly merging information from multimedia descriptors, voice features, and facial patterns. The key contributions of the proposed framework for Facial Expression Recognition and emotion classification can be summarized as follows:

1) The system offers a complete method to capture subtle emotions by integrating three different feature sets: prosodic characteristics, a unique Audio-Visual Descriptor, and Weighted Edge Local Directional Pattern (WELDP). This lessens the reliance on present feature sets.

2) WELDP optimises the encoding process by using seven bits to represent each pixel, hence minimising code repetition.

3) This helps to portray face characteristics more effectively and efficiently, especially when there are weak or warped edges present.

4) The suggested system leverages both individual models by combining the Support Vector Machine and Decision Tree methods for categorization.

5) Combining the output from various models improves the prediction of emotions and yields a categorization result that is more trustworthy.

6) A novel method is provided by the introduction of the Audio-Visual Descriptor based on the Self-Similarity Distance Matrix (SSDM).

7) SSDM computes spatial and temporal distances using facial landmark points, giving priority to face alignment and key frame selection for Audio-Visual input.

8) This results in a succinct yet powerful depiction of emotion in a variety of settings.

The format for the enduring paragraphs is as follows: The relevant work based on various methodologies for diabetes prediction is examined in Section II, and the research gap is identified in Section III. The feature selection and classification process for the proposed method is explained in the Section IV. The outcomes and considerations are covered in Section V; the prospective applications for the future are covered in Section VI.

II. Related Works

The techniques for identifying emotions using multiple interfaces EEG data and multipurpose physiological indicators are the main topic of this extensive literature analysis [14]. The research employs a conventional emotional identification pipeline, looking at different approaches to extracting features like wavelet transformed and nonlinear behaviour, and also decreasing features and machine learning (ML) classification system design methodologies like k-nearest neighbour (KNN), naive Bayesian (NB), support vector machine (SVM), and random forest (RF). The work delves deeper into the complex relationship between various brain regions and mental conditions by analysing EEG patterns that are significantly connected with sentiments. The paper also compares and contrasts machine learning and deep learning methods for emotion identification, highlighting the advantages and disadvantages of each. The study finds that the scope of the collection of features and the data set used for training have a major impact on how well DL models can recognise emotions.

Communication between humans and computers has greatly increased availability of educational resources, data, and the sharing of significant skills on a worldwide scale. The model in [15] offered suggests fusing speech qualities and face emotions to overcome these drawbacks. The technique particularly employs the Speech (Mel Frequency Cepstral Coefficients) and Facial (Maximally Stable Extremal Regions) aspects, which add to an organised and methodical investigation. This methodology presents a more resilient and adaptable emotion identification system, highlighting the possibility of using multifaceted characteristics for enhanced accuracy in a variety of situations in life. Although the bi-modal approach which combines voice and face features has shown to improve accuracy and resilience in the sense of emotion identification network when compared with unimodal. This method, one significant drawback is the requirement for additional expansion.

Two neural network methods in [16] for recognising and categorising objects are included in the suggested flexible design. These approaches are taught separately to facilitate use in real time. AdaBoost cascading models are used for face identification, and then neighbourhood difference features (NDF) are extracted to get localised attractiveness knowledge. This allows for the development of various structures depending on the interactions between surrounding areas. The adaptable structure of the system, which focuses on seven primary facial gestures, enables it to be extended to categorise a wide variety of emotions. In order to manage mis-/false detection, a classifier using random forests with a deep mood indicator has been developed, providing self-determination through sexual orientation and face complexion. The investigation of shape characteristics and face modification brings complexities that can necessitate advanced mathematical methods, which could have an impact on the computing effectiveness of immediate video analysis.

This study in [17] offers a thorough overview of micro-expression evaluation using videos, highlighting the distinctions among broad and tiny movements and using them as a foundation for assessment. The research presents a fresh collection of data, the micro-and-macro emotion warehouse (MMEW), with a greater quantity of video frames and tagged groups of emotions in recognition of the shortcomings of the current micro-expression databases. The writers compare typical techniques uniformly on CAS(ME)2 for finding and on MMEW and SAMM for classification. In the latter section of this investigation, several avenues for further study are discussed, emphasising the dynamic environment and continuous difficulties in the field of video-based micro-expression evaluation. Quick breakthroughs or changes might arise from the multifaceted nature of micro-expression research and the ever-changing nature of technological devices, requiring regular updates exceeding the scope of this paper. To get latest findings in this quickly increasing field of research, investigators should be aware of how the environment is changing and take into account further sources.

By creating an in-depth structure [18] that combines three different classifiers a deep neural network (DNN), a convolutional neural network (CNN), and a recurrent neural network (RNN), this study tackles the difficult job of audio emotion detection. The approach uses segment-level mel-spectrograms (MS), frame-level modest descriptions (LLDs), and utterance-level outcomes of the highest-level statistical algorithms (HSFs) on LLDs to concentrate on categorising detection of four different feelings. Adopting a multifaceted learning approach, the separate versions of LLD-RNN, MS-CNN, and HSF-DNN that are produced are merged to conduct extraction on continual emotion characteristics and categorise defined types of emotions all at once. It is important to take into account the computing requirements of these optimisation techniques, especially if working with huge data sets or apps that operate in real time. This highlights the necessity for efficient and flexible systems in further research.

The promise of blended learning for automated identification of emotions is discussed in this work, with a focus on the relationships and dependencies between sight and aural domains that are still poorly understood. This approach [19] to Multimodal Emotion Recognition Metric Learning (MERML) seeks to simultaneously learn an adequate representational in a space known as latent and a selective score for both techniques, appreciating the distinctive features of each. A Support Vector Machine (SVM) kernel founded on the Radial Basis Function (RBF) effectively applies the learnt metric. The work acknowledges that developing an efficient measure in multi-modal settings is an important aim for a variety of use cases involving machine learning. There are worries over adaptability across different datasets with different channels and emotions as the evaluation's statistics' unique qualities and complexity may have an impact on MERML's efficacy. For assessing MERML's realistic practicality in cases outside of the assessed datasets, more research is necessary to test its effectiveness and versatility in applications that utilise real-time or huge data sets. Even if this research shows better performance, there is still work to be done on the accessibility of the learnt measure and its applicability to other multimodal recognition of emotions activities.

The literature addresses traditional emotional identification pipelines, like subject-independent behavioural characteristics and machine learning classification algorithms, but also acknowledges the field's developing problems. A possible solution is the suggested bi-modal strategy that combines speech and facial clues; nonetheless, the study rightly highlights the need for further research, especially when dealing with a variety of human faces and environments. Moreover, the study of face expression recognition in autonomous vehicles and human-computer interactions highlights the necessity for thorough assessment, particularly with the influence of temporal and geographical factors. The study on micro-expression evaluation emphasizes the dynamic nature of technology in the sector and emphasizes the significance of dataset limits. While highlighting the need for more research and highlighting the difficulties of complicated models, the work on audio emotion recognition presents intriguing multimodal learning approaches. In a similar vein, while Multimodal Emotion Recognition Metric Learning (MERML) research shows good results, more investigation is necessary to ensure that the model is flexible enough to work with different datasets and in real-world circumstances. In summary, these assessments of the literature successfully highlight the deficiencies and intricacies in the field of emotion recognition, offering significant perspectives for further studies.

III. Problem Statement

This comprehensive overview of the literature aims to address the central problem of identifying mental states using different interfaces, with particular focus on EEG results and numerous physiological markers. The research explores conventional emotional identification pipelines, which include smaller feature sizes and classification algorithms such as KNN, NB, SVM, and RF, in addition to methods like wavelet modification and nonlinear behaviour extraction. Through examining EEG patterns linked to affect, the study investigates the complex relationship between mental states and particular brain regions [14]. It also clarifies the benefits and limitations of using machine learning for emotion recognition as opposed to deep learning methods. The study also addresses the difficulties that arise when applying deep learning techniques in practical settings and emphasizes how important feature selection and data accessibility are in deciding how effective deep learning models are. In order to improve accuracy and overcome the shortcomings of current emotion identification methods, the study suggests a bi-modal approach that combines facial and voice signals [15]. It also explores the rapidly developing field of computational learning for emotion recognition and uses neural network techniques for real-time applications, namely facial expression recognition in HCI, autonomous driving, and micro-expression analysis in movies. Moreover, the paper addresses the complexities of auditory emotion recognition and provides an extensive framework with three different categorization techniques for real-time learning and improved accuracy [18].These difficulties are all related to the larger objective of creating more reliable and effective emotion detection techniques, which emphasizes the critical need for additional study to get beyond present barriers and progress the area.

IV. Proposed SVM and Decision Tree for Emotion Recognition

Datasets from reliable sources, like the Surrey Audio-Visual Expressed Emotion (SAVEE), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Ryerson Multimedia Laboratory (RML), are first acquired for the suggested application. The later steps in the process are built upon these datasets. To ensure consistent feature scaling across various datasets, the data goes through a preprocessing step wherein Min-Max normalization is done. Finding pertinent features in the pre-processed data, the feature extraction procedure applies the Local Directional Pattern (LDP) approach. After that, the Fisher criteria—a technique that finds distinguishing features—are applied to feature selection and dimensionality reduction, maximizing the dataset for further examination. Support vector machines (SVMs) and decision trees are used as classifiers to improve classification accuracy. These classifiers use the chosen features to classify the dataset's emotional content. Furthermore, the Self-Similarity Distance Matrix (SSDM) is utilized as a metric to evaluate the degree of similarity across emotional patterns included in the data. The process of integrating the outputs from various classifiers and matrices is called emotion recognition fusion, and it is the culmination of the suggested technique. By utilizing each component's unique capabilities, this fusion method seeks to increase the overall accuracy and dependability of emotion recognition across the datasets. The all-encompassing strategy, which combines preprocessing, feature extraction, classification, and fusion, highlights how reliable and successful the suggested method, is in identifying emotions in audio-visual data. Fig. 1 explains the conceptual Diagram.

A. Dataset Collection

The data collections utilised in the present investigation are from the Ryerson Multimedia Laboratory (RML), Surrey Audio-Visual Expressed Emotion (SAVEE), and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The 1440 audio recordings in the RAVDESS data set are spoken in English by 12 male and 12 female participants. It is composed of eight distinct emotions: fear, fearlessness, rage, calmness, happiness, surprise, disgust, sadness, and neutrality. The 480 videos in the SAVEE data set are narrated in English by four male participants. It is composed of seven distinct emotions: fear, fearless, pleased, astonished, angry, dissatisfied and sad. The 720 videos in the RML data set are voiced in Persian, English, Italian, Urdu, Chinese, and Punjabi [20].

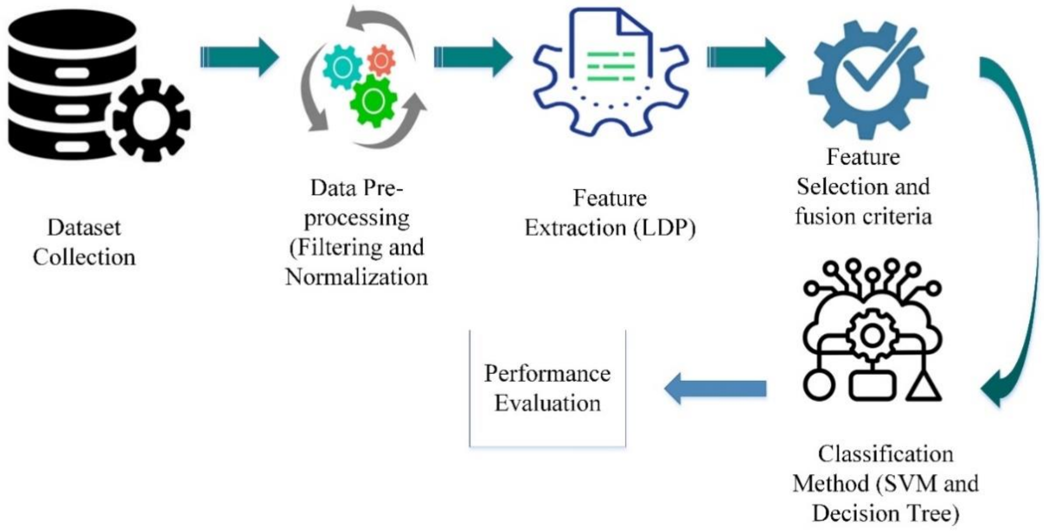
Dataset Collection
Data Pre-processing (Filtering and Normalization
Feature Extraction (LDP)
Feature Selection and fusion criteria
Classification Method (SVM and Decision Tree)
Performance Evaluation

Fig. 1. Overall conceptual diagram.

B. *Data Pre-Processing*

The method of identifying speech and visual disorders is thought to begin with pre-processing. It is used for both pattern extraction and data recognition. Pre-processing techniques are used to modify or amend the speech or visual data, *X(k)*, in order to prepare it for further processing. There is sometimes too much unnecessary and disruptive noise in the transmission of voice or visual data. Traditionally, sound-absorbing cotton or directional microphones were employed to cover up background noise produced by unwanted signals, such as noise from winds and object movement. It is a process for verifying the voice/vision message, *x(k).*If the voice/vision transmission is affected by the surroundings or ambient noise, like *a(k),*this is known as additive disruption. This can be removed using Eq. (1).

$$x(k) = s(k) + a(k) \tag{1}$$

The previous equation yields *s(k),*an exact voice/vision communication. Numerous noise reduction methods can be applied to a noisy speech stream to complete the procedure. The windowing technique: At this stage, the speech or vision information has been split into segments. The windowing functionality, or *w(k),*whose total length is represented by the character l, where l is the audio message's frame length, is used to amplify the signal's frames. Windowing is a type of analysis method where a speech/vision message section in a waveform is multiplied by a time window for a specific form to highlight the signal's intended distinctiveness is shown in Eq. (2).

$$w(k) = 0.54 - 0.46\cos\left(\frac{6.28k}{p-1}\right), \quad 0 \le k \le p-1 \tag{2}$$

Preprocessing must be normalized in order to be categorized. To expedite the learning procedure, the given information has to be normalized. Also, some kind of data normalization might be required to prevent numerical issues like accuracy loss from arithmetic mistakes. Attributes with large starting ranges are likely to dominate an upward descent after first outweighing characteristics with lesser beginning ranges. Since feature space normalization is not applied to the input vectors outwardly, it may be better understood as a kernel perception of preparation than as a specific kind of preprocessing. For instance, in some elements of detection of intrusions statistics, the highest and lowest points in typical and assault are varied by between nine and ten times. Put another way, normalization is a unique kernel mapping method that makes computations easier by converting the data onto a useful plane. The complicated normalization procedure would require an extended period to process because of the massive volume of data points. The chosen Min-Max normalization method is efficient and rapid.

The real information m is transformed linear into the necessary interval max_n, min_n by applying Min-Max Normalization in Eq. (3).

$$m = min_n + (max_n - min_n) * \left(\frac{m - min_u}{max_u - min_u}\right) \tag{3}$$

One benefit of the approach is that it accurately preserves the relationships between the locations in the information. There is no possibility that it could in any way skew the statistics.

C. *Feature Extraction using LDP Method*

When taking into account variations in age, different orientations, and sizes, illuminating effects, and pose variations, regional characteristics outperform global features. Consequently, for AIFR, we have suggested a texture local description. The highly prejudiced local characteristic from the facial components is found by using the suggested descriptor. The variation in the pattern generates a histogram by computing the particular region's pixel differential from the triplet's layout and the relationship among the dual-directional designs. The suggested extracting features method uses the suggested descriptors' local differential pattern [21].

1) Local difference pattern: The appearance and shape of a face evolve with age, making facial recognition harder. To identify a similarity feature that is robust against intra-class variation, research have, consequently, calculated the disparity among pixels of a local region of dimension $a \times a$in an image of dimension $X \times Y$. The suggested feature descriptor covers every local region of a face image.

The set of all the local region variance sequences is represented by Eq. (4),

$$LDP = \left[LDP^{1}_{R^{1}_{u,v}}, LDP^{2}_{R^{2}_{u,v}}, LDP^{3}_{R^{3}_{u,v}}, \ldots..LDP^{Y_i \times Y_j}_{R^{1}_{u,v}}\right] \qquad (4)$$

where, u = 1, 2, 3., X − a +1 and v = 1, 2, 3., Y− a + 1.

For a single local region of $LDP^{1}_{uv}, LDP^{1}_{R^{1}_{u,v}}$is calculated. The pixels $A_{u,v}$ in the local region *LR1 i, j* have intensity values$U_{u,v}$, where *∀u,v = 1, 2, 3, …p.*

$$LR^{1d}_{1,v} = \left[A^{1}_{1,v}, A^{2}_{1,v}, \ldots.A^{1}_{1,v}, ..A^{1*a}_{1,v}, A^{p+1}_{1,v}, A^{p+1}_{2,v}, \ldots..A^{a*a}_{a,v}\right] \quad (5)$$

All of the local region pixel intensity values have been organised in three different formatsrow-wise, column-wise, and orderedto compute the local difference pattern of a face image. The calculation of the difference structure is then performed by transposing the resulting data. The pixels in Eq. (5) have been arranged row-wise and assigned to the 1-dimensional array$A^{1*a}_{1,v}$. The intensities of the correlating pixels are numbered as I and organised in the identical order, with an increasing order of magnitude as the superscript as mentioned in Eq. (6).

$$LR^{1^{1d}}_{u,v} = [U^{1}_{u.v}, U^{2}_{u.v}, U^{3}_{u.v} \ldots.U^{a*a}_{u.v}] \qquad (6)$$

where, *u,v = 1,2,…a*. The subsequent Eq. (7) is used to compute the row-wise, column-wise, and ordered pattern variance of the intensity of the pixels of local area$k = a \times a$.

$$\gamma^{k}_{u,v+1} = \left[\gamma^{k-1}_{u,v+1} - \gamma^{k-1}_{u,v}\right] \qquad \forall u = 1, \forall v = k - 1 \quad (7)$$

Ultimately, the aforementioned formulas are used for calculating the k^{th}difference pattern. Finding the sum of the directed difference pattern, row-wise difference pattern, and column-wise difference pattern represented as $\gamma^{k}_{u,v}$ will yield the final absolute value of a local region.

The distinction arrangement in a single local region is calculated using the above equation; similarly, all variance patterns are obtained to form the final LDP feature vector, which is used to find the histogram. Eq. (8) displays the difference pattern's final feature vector:

$$LDP = \left[LDP^{1}_{LR^{1}_{u,v}}, LDP^{2}_{LR^{2}_{u,v}}, LDP^{3}_{LR^{3}_{u,v}}, \ldots..LDP^{v}_{LR^{1}_{u,v}}\right] \quad (8)$$

A feature vector Local Difference Pattern (LDP) of a face image contains all of the computed values. To generate a unique code for every neighbourhood, the process entails thresholding the intensity differences and assigning binary values. The resulting patterns draw attention to differences in intensity by highlighting textural details such as corners or edges. For tasks like texture analysis, object detection, and facial recognition, LDP is especially helpful because it efficiently encodes local information that can be useful for differentiating between various regions in an image. All things considered; Local Difference Pattern is a reliable technique for obtaining discriminative features that can improve the efficiency of a range of computer vision applications.

2) Spectral features: The present investigation incorporates MFCC, LTAS, and three spectral characteristics formants. Formants are an illustration of the resonant that occurs in the vocal cords at the peak of high intensity. Since formants change with emotion, they are often used in speech emotion recognition. Then measure the MFCC from a quadratic Me-scale to investigate the significance of low-frequency variables in comparison to high-frequency elements. Since they are highly responsive to variations in sounds at lower ranges, they are often used in voice and recognition of speech mechanisms as they mimic the way people's hearing systems adjust for tone and the exponential signal power ratio of vocal parts of speech signals. Furthermore, Long-Term Average Spectrum (LTAS) has less computational demand than MFCC. The final three formants, the overall mean of the LTAS, the average of the 12 MFCCs, and the ranges, greatest, and lowest constitute the characteristics of segments [22]. Fig. 2 shows the workflow of Mel-frequency cepstral coefficients (MFCC).

The feature-extracting method known as MFCC collects both non-linear as well as linear features, which are necessary for speaker identification. The frequency spectrum employed by the MFCC adjusts the frequency exponentially for frequencies above and below 1 kHz. With MFCC, the crucial component that constitutes sound transmissions may be recorded. The acronym for complex cepstral coefficients is MFCC. Since the MFCC provides both time and frequency information about the signal, it is more beneficial for feature extraction. MFCCs have found widespread application in voice recognition due to their ability to effectively handle dynamic features of the audio data while capturing all non-linear and linearity qualities. Since the sounds have both non-linear as well as linear properties, MFCC is a useful technique for feature extraction. MFCC is a particularly often employed feature in contemporary speech verification/identification methods, according to a number of studies. These factors make MFCC a popular choice for speech recognition systems:

- The discrete cosine transformation (DCT) effectively makes the cepstral features orthogonal.
- Noise from stationary channels is eliminated by subtracting the cepstral average.
- MFCC is less vulnerable to additive noise than other feature extraction techniques like linear prediction of cepstral coefficients.

The following are the methods that MFCC employs to extract features: Initially signal pre-processing is applied to a voice message. It applies pre-emphasis filtering to equalize the exact dimensions. A Hamming is the Window that has been attached to each block to mitigate the edge impacts caused by the window cutting. A discrete cosine transformation is applied to the processed signal, and it is then softened using a series of triangle filters separated along a Me1 Scale.

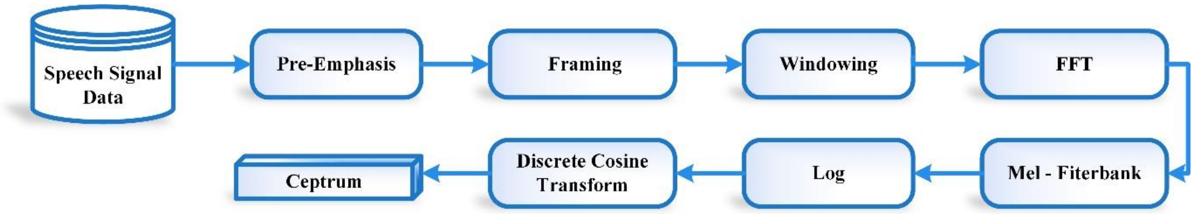

Fig. 2. MFCC workflow.

In audio and voice audio processing, MFCCs are a frequently used feature extraction method. The Fast Fourier Transform (FFT) is used to transform the data into the domain of frequencies, and filter bank processing is then applied to replicate human aural perception. After applying logarithmic compression to the filter bank energies, the MFCCs which capture crucial spectral properties for tasks like sound classification and speech recognition are obtained by the discrete cosine transform (DCT).

D. Feature Selection and Dimensionality Reduction

The purpose of selecting features is to ascertain the characteristics' relative relevance. Following the process of extraction aspects of the input voice signal, only significant characteristics are selected, with the other features eliminated. The characteristics' relevance is calculated at this stage. In this paper, they implement the feature selection procedure using correlational evaluation. In addition, they employ a linear discriminatory analysis approach for feature reducing dimensionality called the Fisher criteria. Initially, Euclidean distance investigation, partial correlation estimation, and vicariate correlation assessment are used to choose the features. The final features with decreased dimensions are then obtained by applying the Fisher criteria to the consequent aspects that were acquired [22].

1) Correlation analysis: The Euclidean distance analysis is completed at this point, and all of the characteristics are grouped together. The correlation between the attributes in each category is then determined by using a modified study of correlation on each group. The final feature set is then determined by evaluating the resulting features for Spearman rank correlation (SRC) assessment. Following the start of the correlation investigation, the resulting emotional traits become increasingly apparent.

2) Euclidean distance analysis: An individual's emotions may be described by a variety of qualities, but it might be difficult to apply them concurrently when conducting emotion recognition. As such, it is crucial to determine which traits are fundamental and have a significant impact on management and emotions. Each characteristic is first examined to see how it relates to other characteristics; those characteristics are then categorised according to the results of this analysis. In this case, the features are grouped using distance analysis. The Euclidean distance shows the true distance between two points in an aspect set of n dimensions. Eq. (9) expresses the Euclidean distance between two points, (X_1,Y_1) and (X_2,Y_2).

$$E = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{9}$$

where, characteristics that have lower *E*scores are clustered into a single group. E indicates the Euclidean distance. $X=X_1,X_2, ...,X_n$ *and* $Y=Y_1,Y_2, ..., Y_n$are locations in a space with n dimensions. Equivalent characteristics are chosen from the closest proximity characteristics.

3) Partial correlation analysis: Since many emotional components share characteristics with a state of feeling, it may be difficult to ascertain how features influence the emotional state. Before examining the link between characteristics and associated feelings, it is necessary to exclude or regulate the features that negatively impact the other aspects. One may refer to this type of study as net correlational analysis or partial correlation assessment. This type of research uses the linear relationship across two features to identify how one affects the other. $X= \{X_1,X_2, ...,X_n\}$ is the set of independent factors; the partial correlation between these variables is calculated using Eq. (10)

$$r = (\rho^{ij})_{n\times n} = \begin{bmatrix} \rho^{11} & \cdots & \rho^{1n} \\ \vdots & \ddots & \vdots \\ \rho^{n1} & \cdots & \rho^{nn} \end{bmatrix} \tag{10}$$

Eq. (11) is used to obtain the inversion for the Matrix mentioned above.

$$r^{-1} = (\lambda^{ij})_{n\times n} = \begin{bmatrix} \lambda^{11} & \cdots & \lambda^{1n} \\ \vdots & \ddots & \vdots \\ \lambda^{n1} & \cdots & \lambda^{nn} \end{bmatrix} \tag{11}$$

Eq. (12) is used to determine the partial relationship among the two variables.

$$Y^{ij} = \frac{-\lambda^{ij}}{\sqrt{\lambda^{ii}}\sqrt{\lambda^{jj}}} \tag{12}$$

The relation between the two separate variables is defined by the coefficient. It subtly illustrates their dependence and the need for selecting or elimination.

4) Nonlinear correlation analysis: There are multiple techniques available when calculating the correlation coefficient between both variables which are a Spearman Rank Correlation (SRC), a Kendall Coefficient of Concordance (KCC), and a Pearson Product Moment Linear Correlation Coefficient (PLCC). However, the characteristics that were taken from all frames inside every time frame are distinct, completely sorted variables, and we used the SRC approach to determine their rank. One kind of index that examines the statistically significant relationship between two variables under a linear function is the SRC Coefficient. The SRC Coefficient is either +1 or - 1 for variables that are strictly monotonic to one another; these variables are referred to be full spearman correlations. Let $X= \{X_1,X_2, ..., X_n\}$, and $Y=\{Y_1,Y_2, ..., Y_n\}$be two variables. Using Eq. (13), the SRC coefficient for each is determined.

$$\rho^s = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \qquad (13)$$

In the X variable, Xirepresents i-th characteristic, this is whereas Yi represents*i-th* feature in the Y variable. The Mean of X and Y are denoted by $\bar{X}$ and $\bar{Y}$, respectively. A not-parametric correlation value is typically used to calculate the SRC coefficient. When knowing the combined probability distribution of X and Y samples, the SRC Coefficient calculates precisely the distribution of the two observations. Between X and Y, the SRC Coefficient is present up through the monotonic connection. The SRC is not the same as the PLCC, which is only dependent on linearity characteristics.

5) Fisher criterion: The multidimensionality of the feature set presents a number of challenges in the development of statistical algorithms for recognised patterns purposes. Low dimensionality methods can provide optimal performance with minimal computing burden. More relevant characteristics are acquired at the feature identification stage, and they undergo a transformation into lower-dimensional space with very little data loss. The loss of data is the main problem of reduction in dimensionality. Therefore, we use the Fisher Criterion, which establishes the concept of linear relation-based reduction in dimensions, to provide an ideal set of attributes with low dimensionality space. Another well-liked technique for reducing complexity is PCA, yet it is unable to separate differentiates between low from very high multidimensional emotional traits. Eq. (14) is used to compute the Fisher Criterion analytically.

$$\lambda^F = \frac{\sigma^B}{\sigma^W} \qquad (14)$$

where, σ^W denotes the variation inside the class, σ^B denotes the variance among classes, and λ^F stands for Fisher's rate for characteristics. Eq. (15) definesσ^B.

$$\sigma^B = \sum_{c=1}^{N}(E^c - \bar{E})(E^c - \bar{E})^T \qquad (15)$$

where, as specified in Eq. (16),$\bar{E}$ is the average of the whole set of data.

$$\bar{E} = \frac{1}{m}\sum_{i=1}^{m} X_i \qquad (16)$$

Moreover,E^c which is specified in Eq. (17), is the sample's average for *ith* Emotions classes.

$$E^c = \frac{1}{N_p}\sum_{X\epsilon E^c} X_i \qquad (17)$$

The whole quantity of feelings is denoted by the value M in Eq. (16) and the total amount of instances in the speech's emotional signals is denoted by the expression in Eq. (17). Likewise, it has a mathematical definition found in Eq. (18).

$$\sigma^w = \sum_{c=1}^{N}\sum_{i=1}^{N_p}(X_i - E^c)(X_i - E^c)^T \qquad (18)$$

They next used decrease in dimensionality to the shape of the distribution vector, σ^w in order to eliminate extraneous features while keeping the crucial data intact.

E. SVM, Decision Tree Classifier

The total amount of characteristics defines N, which is utilized to find the hyperplane in the space with N dimensions using SVM. The hyperplane facilitates the information point classification procedure. The greatest distance on the plane among the different categories should be used. Non-linear (RBF) kernels are employed in this study to classify support vectors. Using Eq. (4), the kernel aids in obtaining the hyperplane for identifying various classes in Eq. (19):

$$K(X^i, X^j) = e^{-\gamma\|X^i - X^j\|^2} \qquad (19)$$

where, the squared Euclidean spacing among the two input information vectors, X^i and $X^j$is represented as $\left\|X^i - X^j\right\|^2$. The incorrect classification rate, C, for the study described in this article, has been fixed at 2. The percentage of inaccurate classifications made by the model that was trained is known as the misperception rate. In order to get the most significant margins between the classes and to signal a smaller erroneous bound, a minimum amount was used [23]. Fig. 3 shows the structure of SVM.

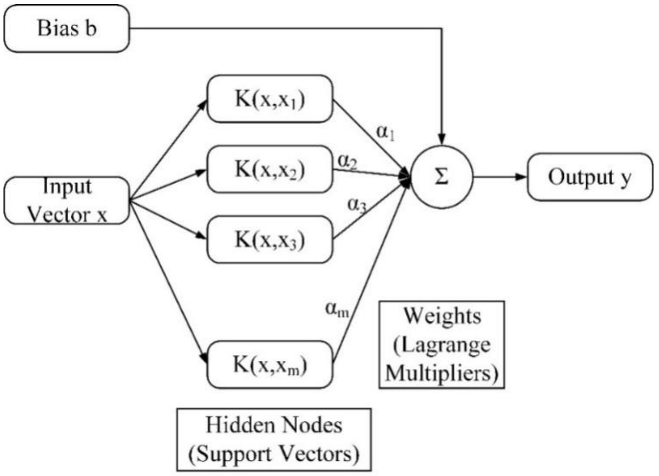

Fig. 3. Structure of SVM.

An optimal splitting of the input characteristics is generated to choose the nodes of a tree that makes up the DT classifier. The greatest information gain (IG) is produced by the separation of the data itself and the tree roots. The tree pruning had been configured with a maximum cutting depth of 8 in order to prevent overfitting of the simulation. Obtaining the parental node's IG (20) was applied.

$$IG(D^p) = ID^p - \frac{N_{left}}{N^p} ID_{left} - \frac{N_{right}}{N^p} ID_{right} \qquad (20)$$

where, an array comprising the parent, left, and right datasets is represented by D^p, D_{left} , and D_{right}. In this investigation, I, the value of entropy, was employed to determine the separated entropy criterion's efficiency. Eq. (21) was used to compute the entropy.

$$I = -\sum_i p^i . log_2 p^i \qquad (21)$$

The value that is targeted i'sprobability is represented byp^i. Eq. (22) was utilized in addition to determining the categorization error.

$$DTmodel^{error} = 1 - \max(p^i) \tag{22}$$

F. Self-Similarity Distance Matrix (SSDM)

They represent every movie as a collection filled with regional SSM descriptors H and use recently effective bags-of-features techniques to identify emotion events. Next, they train and categorize examples that represent action classes using SVM.

They create visual text histograms and utilize them as a source of inputs for training using SVM and classifications. Using 10,000 local SSM descriptors (h) divided into k = 1000 clustered from the training collection, a visual language is created. The graphical representation of visual phrases is then calculated for every image in the sequence, and every characteristic is then paired with the vocabulary word that is the closest to it (although they utilize the Euclidean distances). They use the χ 2 kernels for training, not linear SVMs, and use a one-versus-all strategy for the classification of multiple classes [24].

They give n-fold cross-validation findings for each of the recognition investigations in the following section and ensure that a single person's behaviours are not seen in both the training and test sets at the same time.

G. Emotion Recognition Fusion

The methods mentioned above deal with single-frame and 400 ms sound segment prediction. The frame-based forecast was converted to a video-based projection so that the outcomes could be compared with human users. Similarly, to how the audio recordings are divided into manageable chunks, the outcomes were combined to provide one recorded prediction. The following procedure was followed in order to make a single prediction: Each of the six scores for audio and video predictions relates to the expected accuracy for a certain class. Since all probabilities add up to one, the label that is predicted is represented by the value with the greatest sum [25]. To obtain the final forecasting, the individual probabilities are added together and normalized. To get a single forecast for an instance file (audio + video) in a similar fashion, the earlier indicated technique is required. Since the six probabilities form the basis of the audio- and video-predicted labels, these can be easily compared, a process known as decision-level fusion.

V. Result and Discussion

The proposed method's performance evaluation shows notable improvements in the accuracy of emotion recognition. Based on thorough experimental validation using the SAVEE, RAVDESS and RML datasets, the combined strategy shows significant gains over current techniques. A comprehensive feature set that enhances classification is demonstrated by the combination of speech attributes and Weighted Edge Local Directional Patterns for facial pattern extraction.

Metrics for emotion recognition accuracy are displayed in the Table I for a range of emotional categories in three different datasets: SAVEE, RAVDESS, and RML. The reliability for the method of classification is further highlighted by the use of Decision Tree and Support Vector Machine algorithms. The approach gains a valuable dimension with the addition of the novel's Audio-Visual Descriptor, which focuses upon facial alignment as well as selection. This leads to an improvement in emotion recognition performance. The cohesive integration of speech characteristics, multimedia descriptors, and facial patterns enhances the method's adaptability and resilience in identifying feelings within a variety of methods. Based on the percentage accuracy of detecting seven different emotions angry, calm, disgusted, fearful, happy, sad, and surprised each dataset is evaluated. In addition, the SAVEE dataset scored 75% accuracy within the angry category, the RAVDESS dataset obtained 90% accuracy, and the RML dataset showed a high accuracy of 93% accuracy. To identify the remaining emotions within every dataset, the table also gives accuracy percentages. These metrics are useful for comparing how well emotion detection models trained regarding the different datasets perform, showing the different levels of success in correctly classifying different emotional states.

Fig. 4 shows the performance of emotion recognition using accuracy measures for different emotions across three different datasets: SAVEE, RAVDESS, and RML. Particularly, the RML dataset performs well in the angry category with an accuracy of 93%, outperforming both SAVEE (75%) and RAVDESS (90%). The SAVEE dataset is better at identifying Fearful emotions (71%) than the RAVDESS dataset is at identifying Calm emotions (95%). Accuracy in the happy category is comparatively similar, using RAVDESS and RML receiving scores of 86% and 88%, correspondingly. The RML dataset consistently scores well within the disgusted and surprised categories, achieving accuracies of 87% and 88%. The significance of choosing relevant datasets over training models customized to particular emotional states is highlighted by these results, which highlight the dataset-specific details in emotion recognition.

TABLE I. PERFORMANCE OF THE DATASET

Dataset Metrics	SAVEE	RAVDESS	RML
Angry	75	90	93
Calm	73	95	90
Disgusted	70	88	87
Fearful	71	88	84
Happy	69	86	88
Sad	70	83	85
Surprised	68	87	88

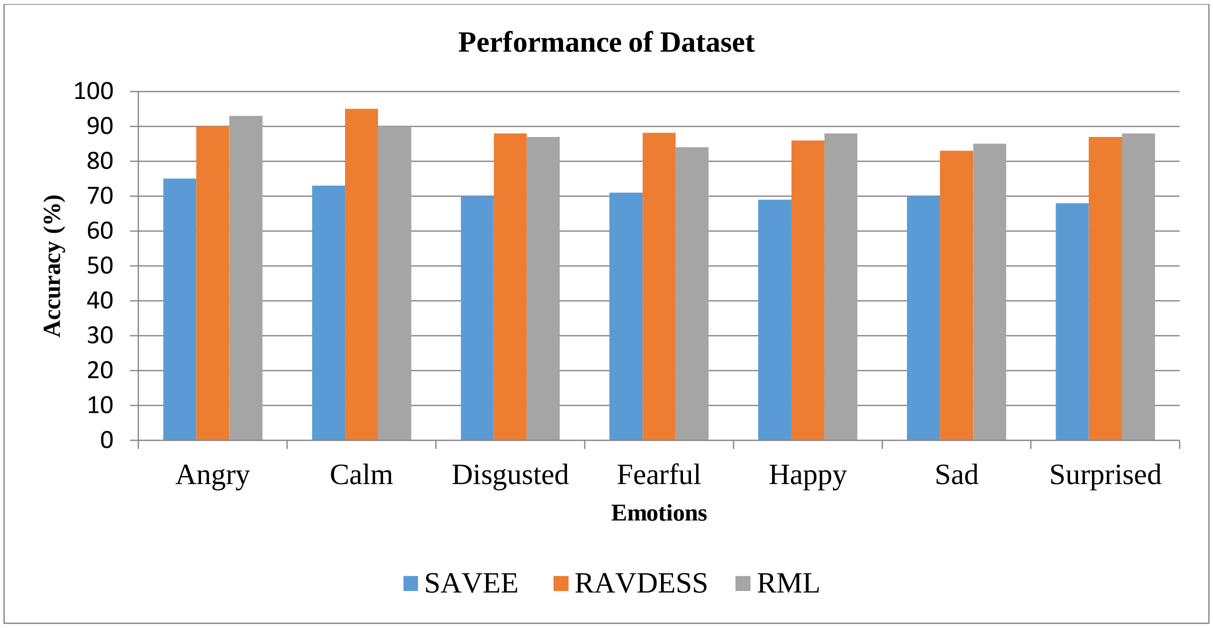

Fig. 4. Graphical representation for performance of the dataset

CNN-LSTM, CNN with Class Activation Mapping, Transformer with Self-attention, the suggested SVM and Decision Tree approach, and other classification techniques are all thoroughly compared in Table II that is presented. With an astounding accuracy of 98.2%, the suggested SVM and Decision Tree technique notably beats its competitors, demonstrating its efficacy in correctly classifying data. The robustness and reliability of the suggested strategy are shown by the precision, recall, and F1-Measure scores of 98.6%, 97.9%, and 98.3%, respectively. With improvements in classification accuracy over the most advanced techniques, this higher performance establishes the suggested SVM plus Decision Tree approach as a highly competitive solution and makes it a viable option for applications needing accurate and dependable classification.

TABLE II. PERFORMANCE COMPARISON WITH EXISTING AND PROPOSED METHOD

Classification method	Accuracy of Classification Method	Precision	Recall	F1-Measure
CNN-LSTM [26]	80%	83.2%	72%	85%
CNN and Class Activation Mapping [27]	90%	93%	90.5%	88%
Transformer and Self-attention [28]	97%	97.8%	98.1%	98%
Proposed SVM and Decision Tree	98.2%	98.6%	97.9%	98.3%

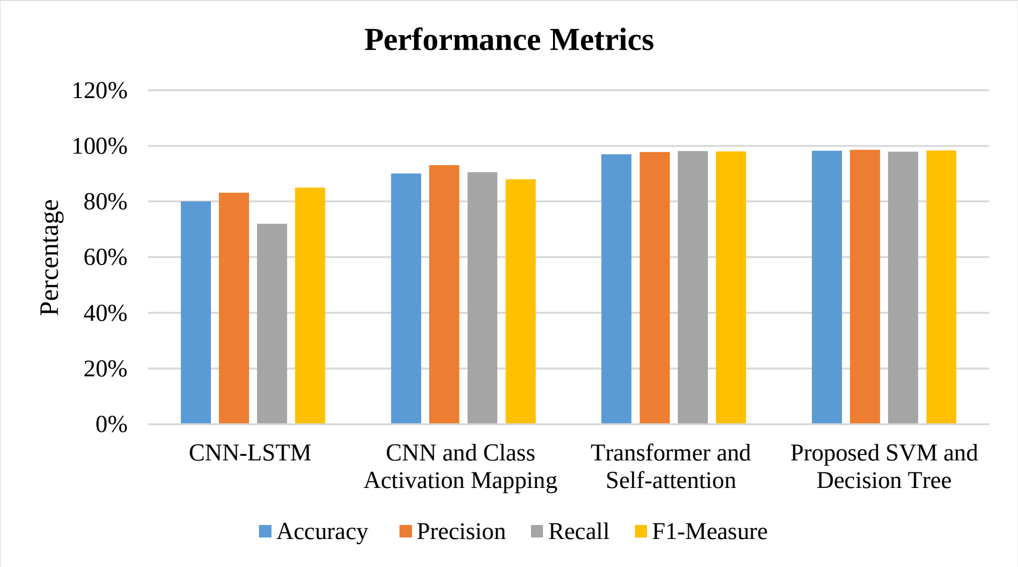

Fig. 5. Graphical representation of performance comparison with existing and proposed method.

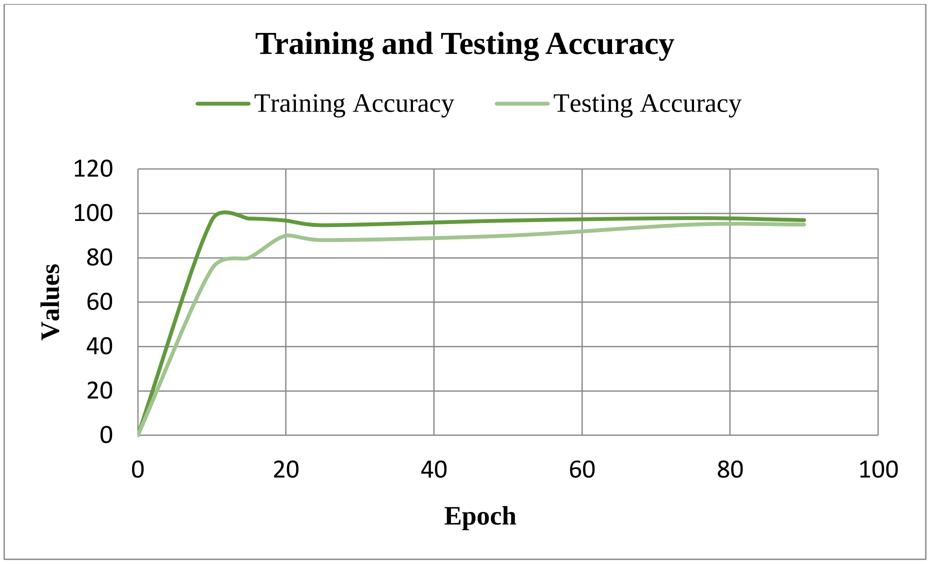

Fig. 6. Graphical depiction for training and testing accuracy of proposed SVM and proposed model.

The performance measures of several classification techniques are shown in the Fig. 5 in terms regarding percentage accuracy. With an accuracy of 80%, the CNN-LSTM model demonstrated its ability to combine long short-term memory (LSTM) networks and convolutional neural networks (CNNs) for classification tasks. A higher accuracy of 90% was shown by the CNN and Class Activation Mapping approach, demonstrating the efficiency of convolutional neural networks enhanced about class activation mapping over localization. With an astounding accuracy of 97%, the Transformer and Self-attention framework demonstrated the effectiveness of self-attention mechanisms in identifying complex patterns. Support vector machines (SVMs) and decision trees work well together to provide accurate classification in the given context. This is demonstrated by the proposed SVM and Decision Tree ensemble, which performed better than other approaches and achieved an impressive accuracy of 98.2%.

Fig. 6 shows the training accuracy graph, within the framework of the proposed SVM and decision tree model, shows how well the model classifies emotions using the training dataset throughout its training iterations, or epochs. This graph shows how the model is learning and if it is becoming more accurate or if the training data may be overfitting. However, the effectiveness of the model on a separate, untested dataset that was not used for training is depicted on the testing accuracy graph. This graph sheds light on how well the model can use its knowledge of emotion recognition to situations outside of the training set. A high testing accuracy shows the model's competence in consistently identifying emotions in a variety of real-world scenarios, and those is essential for improving emotion recognition. A substantial training accuracy suggests that the framework successfully acquired from the training data.

The suggestedSVM and decision tree model demonstrates how the model's loss function changes over the training and evaluation stages in a graphical representation for training and testing loss is shown in Fig. 7. The training loss graph illustrates where the loss, a measure of the discrepancy between expected and actual values, varies throughout the course of the model's training epochs or iterations. The framework learns from and adjusting to the training data when there is a decreasing training loss.The model repeatedly adjusts its variables during training epochs, and the evolution of the loss values which quantify the difference among the predictions made by the model and the actual target values is shown in the training loss curve. The training loss shows that the model is becoming better at fitting the training data. On the other hand, the testing loss curve shows how well the model performs on a test dataset that has not yet been seen, providing information about how well it can generalize and generate correct predictions in practical situations. When faced with new, untested data, the model's capacity to generalize well and provide a lower error rate is demonstrated by the testing loss value that decreases.

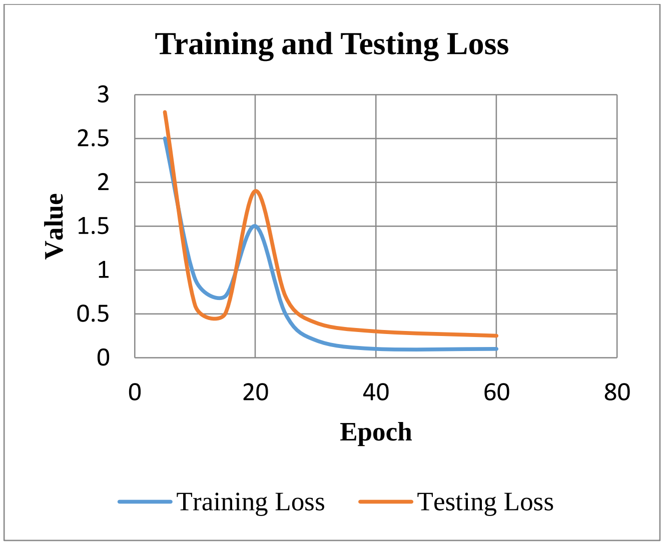

Fig. 7. Graphical depiction for training and testing loss of proposed SVM and decision tree model.

A. Discussion

The outcomes demonstrate the significant progress made by the suggested emotion recognition approach, particularly in contrast to other methods now in use. Using three different datasets, including SAVEE, RAVDESS, and RML, the method's performance is methodically assessed over a range of emotional categories. The increased accuracy of the method's emotion recognition can be attributed to the addition of the Audio-Visual Descriptor, which highlights face alignment and selection. The thorough measurements offer a thorough grasp of how well the approach recognizes seven distinct emotions throughout the datasets. the performance across datasets and emotions, highlighting the subtleties unique to each dataset in the identification of emotions. The suggested SVM and Decision Tree method's exceptional accuracy of 98.2% can be seen when compared to advanced techniques[26] [27] [28]. Its robustness and reliability are further highlighted by the precision, recall, and F1-Measure scores. The suggested approach performs better than CNN-LSTM, CNN with Class Activation Mapping, and Transformer with Self-attention, as demonstrated by the graphical representations. In addition to performing better than existing techniques, the suggested strategy attains a greater accuracy of 98.2%.

The accuracy graphs for training and testing provide information on how the model learns and how well it generalizes to new data. A high testing accuracy shows that the model is useful outside of the training set and demonstrates its competency in real-world circumstances. In addition, the model's learning dynamics are demonstrated by the training and testing loss curves, where lowering loss values denote better generalization and data fitting. The suggested technique for identifying emotions uses a combination of speech characteristics and Weighted Edge Local Directional Patterns, and it shows excellent accuracy and resilience in a range of datasets and emotional classifications. Because of its increased versatility, the Audio-Visual Descriptor presents a viable option for applications that demand accurate and dependable emotion recognition.

This study has certain limitations, despite the notable advancements and gains shown in facial expression analysis and emotion recognition. The suggested model's dependence on edge detection might provide difficulties in situations with different illumination conditions, which could affect the expression assessment's correctness. The model's efficacy can vary depending on the context, and more research may be necessary before applying it to a wider range of real-world settings. The study may have limited cross-platform compatibility and interoperability with other programming languages or frameworks due to its exclusive usage of Python tools for implementation. Further research and development may be needed in the model's ability to handle complex or delicate emotional expressions.

VI. Conclusion and Future Work

In order to overcome difficulties with emotion classification and facial expression identification, this research offers a fresh and practical paradigm. A thorough method for capturing minor emotional cues is demonstrated by the integration of three different feature sets: prosodic characteristics, a new audio-visual descriptor, and Local Differential Pattern (LDP). Specifically designed to improve face feature extraction, the LDP approach is useful for reducing distortion and unreliability associated with edges in facial images. The proposed Audio-Visual Descriptor uses a Self-Similarity Distance Matrix (SSDM) to give a concise description of emotion while giving priority to key frame selection and facial alignment. Experimentation validation using SAVEE, RAVDESS, and RML datasets demonstrates notable advances over existing approaches, highlighting the usefulness of the suggested framework in correctly categorizing emotions and facial expressions in a range of contexts. There are numerous directions for more research and development. More advancement in classification accuracy may be possible by enhancing the suggested framework by adding deep learning architectures, evaluating alternative machine learning algorithms, and examining sophisticated feature extraction methods. Increasing the size and diversity of datasets included in the experimental validation process would improve our comprehension of the generalizability of the model. For practical applications, addressing real-world issues like varied lighting and dynamic face expressions may be essential. It would also be beneficial to investigate how the framework is implemented in real-time settings and evaluate how well it works in erratic environments. In conclusion, the proposed framework raises the bar for the field of emotion identification technology and provides a strong basis for future research projects. Regarding theoretical ramifications, this study makes a substantial contribution to the field by presenting a novel framework that integrates a variety of feature sets, giving emotion categorization a more sophisticated and reliable method. The limitations of current approaches are addressed by the integration of prosodic features, audio-visual descriptors, and LDP, creating opportunities for the exploration of richer emotional cues. From a practical standpoint, the suggested framework has a benefit in that it can reliably categorize emotions and facial expressions in a range of situations. The model has been shown to outperform existing approaches in a number of real-world applications, including affective computing, mental health assessment, and human-computer interaction. These advantages are especially evident in datasets like SAVEE, RAVDESS, and RML. It is essential to recognize the constraints of this research. The accuracy of expression assessment may be affected by the dependence on edge detection algorithms in situations when illumination conditions are changeable. More research and validation are needed to fully understand how the model works in different real-world contexts and how context-specific it is.The suggested framework's potential can be increased for further studies by incorporating deep learning architectures and investigating different methods. A more thorough grasp of the model's capabilities and constraints will also result from extending experimental validation to bigger datasets and tackling real-world issues. Important objectives for future research include investigating real-time implementation and assessing performance in dynamic situations.The foundation for developing emotion recognition technology is laid by this research; however, further investigation and improvement are necessary to ensure the technology's advancement and usefulness.

REFERENCES

[1] W. Z. Khan, Y. Xiang, M. Y. Aalsalem, and Q. Arshad, “Mobile Phone Sensing Systems: A Survey,” IEEE Commun. Surv. Tutor., vol. 15, no. 1, pp. 402–427, 2013, doi: 10.1109/SURV.2012.031412.00077.

[2] R. E. Jack and P. G. Schyns, “The Human Face as a Dynamic Tool for Social Communication,” Curr. Biol., vol. 25, no. 14, pp. R621–R634, Jul. 2015, doi: 10.1016/j.cub.2015.05.052.

[3] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PLOS ONE, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[4] “A review on sentiment analysis and emotion detection from text | SpringerLink.” Accessed: Nov. 23, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s13278-021-00776-6.

[5] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, “Survey on Emotional Body Gesture Recognition.” arXiv, Jan. 23, 2018. Accessed: Nov. 23, 2023. [Online]. Available: http://arxiv.org/abs/1801.07481.

[6] “Frontiers | SlimMe, a Chatbot With Artificial Empathy for Personal Weight Management: System Design and Finding.” Accessed: Nov. 23, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnut.2022.870775/full.

[7] F. González Hernández, R. Zatarain Cabada, M. Barron Estrada, and H. Rodriguez Rangel, “Recognition of learning-centered emotions using a convolutional neural network,” J. Intell. Fuzzy Syst., vol. 34, pp. 3325–3336, May 2018, doi: 10.3233/JIFS-169514.

[8] S. Shah, H. Ghomeshi, E. Vakaj, E. Cooper, and S. Fouad, “A review of natural language processing in contact centre automation,” Pattern Anal. Appl., vol. 26, no. 3, pp. 823–846, Aug. 2023, doi: 10.1007/s10044-023-01182-8.

[9] S. Shayaa et al., “Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges,” IEEE Access, vol. 6, pp. 37807–37827, 2018, doi: 10.1109/ACCESS.2018.2851311.

[10] F. H. Wilhelm and P. Grossman, “Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment,” Biol. Psychol., vol. 84, no. 3, pp. 552–569, Jul. 2010, doi: 10.1016/j.biopsycho.2010.01.017.

[11] “Experiences of black and minority ethnic (BME) students in higher education: applying self-determination theory to understand the BME attainment gap: Studies in Higher Education: Vol 46, No 3.” Accessed: Nov. 23, 2023. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/03075079.2019.1643305.

[12] “Automatic, Dimensional and Continuous Emotion Recognition | International Journal of Synthetic Emotions.” Accessed: Nov. 23, 2023. [Online]. Available: https://dl.acm.org/doi/abs/10.4018/jse.2010101605.

[13] S. Pearson, “Privacy, Security and Trust in Cloud Computing,” in Privacy and Security for Cloud Computing, S. Pearson and G. Yee, Eds., in Computer Communications and Networks. , London: Springer, 2013, pp. 3–42. doi: 10.1007/978-1-4471-4189-1_1.

[14] J. Zhang, Z. Yin, P. Chen, and S. Nichele, “Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review,” Inf. Fusion, vol. 59, pp. 103–126, 2020.

[15] K. Prasada Rao, M. V. P. Chandra Sekhara Rao, and N. Hemanth Chowdary, “An integrated approach to emotion recognition and gender classification,” J. Vis. Commun. Image Represent., vol. 60, pp. 339–345, Apr. 2019, doi: 10.1016/j.jvcir.2019.03.002.

[16] A. Alreshidi and M. Ullah, “Facial Emotion Recognition Using Hybrid Features,” Informatics, vol. 7, no. 1, Art. no. 1, Mar. 2020, doi: 10.3390/informatics7010006.

[17] X. Ben et al., “Video-based facial micro-expression analysis: A survey of datasets, features and algorithms,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5826–5846, 2021.

[18] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, “Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN,” Speech Commun., vol. 120, pp. 11–19, Jun. 2020, doi: 10.1016/j.specom.2020.03.005.

[19] E. Ghaleb, M. Popa, and S. Asteriadis, “Metric Learning Based Multimodal Audio-visual Emotion Recognition,” IEEE Multimed., pp. 1–1, 2019, doi: 10.1109/MMUL.2019.2960219.

[20] O. ATİLA and A. ŞENGÜR, “Automatic Speech Emotion Recognition Using Machine Learning and Iterative Neighborhood Component Analysis,” 2021.

[21] R. K. Tripathi and A. S. Jalal, “Novel local feature extraction for age invariant face recognition,” Expert Syst. Appl., vol. 175, p. 114786, 2021.

[22] K. Ramyasree and C. S. Kumar, “Multi-Attribute Feature Extraction and Selection for Emotion Recognition from Speech Through Machine Learning,” Trait. Signal, vol. 40, no. 1, p. 265, 2023.

[23] C. M. T. Khan, N. A. Ab Aziz, J. E. Raja, S. W. B. Nawawi, and P. Rani, “Evaluation of machine learning algorithms for emotions recognition using electrocardiogram,” Emerg. Sci. J., vol. 7, no. 1, pp. 147–161, 2022.

[24] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez, “Cross-view action recognition from temporal self-similarities,” in Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10, Springer, 2008, pp. 293–306.

[25] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, “Audiovisual emotion recognition in wild,” Mach. Vis. Appl., vol. 30, no. 5, pp. 975–985, 2019.

[26] E. Ryumina, D. Dresvyanskiy, and A. Karpov, “In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study,” Neurocomputing, vol. 514, pp. 435–450, 2022.

[27] C. Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” J. Mach. Learn. Res., vol. 21, no. 1, pp. 5485–5551, 2020.

[28] Y. Zhang, C. Wang, X. Ling, and W. Deng, “Learn from all: Erasing attention consistency for noisy label facial expression recognition,” in European Conference on Computer Vision, Springer, 2022, pp. 418–434.