DeepEmoVision: Unveiling Emotion Dynamics in Video Through Deep Learning Algorithms

Prathwini[1], Prathyakshini[2]*
Department of Master of Computer Applications, NMAM Institute of Technology, NITTE (Deemed to be University), India[1]
Department of Information Science and Engineering, NMAM Institute of Technology, NITTE (Deemed to be University), India[2]

***Abstract*—Emotion detection from videos plays a pivotal role in understanding human behavior and interaction. This study delves into a cutting-edge method that leverages Recurrent Neural Networks (RNN), Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Convolutional Neural Networks (CNN) and to precisely detect emotions exhibited in video content, holding significant importance in comprehending human behavior and interactions. The devised approach entails a multi-phase procedure: initially, employing CNN-based feature extraction to isolate facial expressions and pertinent visual cues by extracting and pre-processing video frames. These extracted features capture intricate patterns and spatial information crucial for discerning emotions. The results of the trials show that CNN, SVM, KNN, and RNN have promising performance, highlighting their potential. Among the other machine learning models, RNN has attained a 95% accuracy rate in recognizing and classifying emotions in video information. This combination of approaches provides a thorough plan for identifying emotions in dynamic visual material in real time.**

Keywords*—Emotion detection; video analysis; Recurrent Neural Networks (RNN); Support Vector Machines (SVM); K-Nearest Neighbours (KNN); Convolutional Neural Networks (CNN); facial expression recognition; machine learning

I. Introduction

The foundation of social dynamics and interpersonal communication is an understanding of human emotions. Human-computer interaction, affective computing, and psychology all heavily rely on the detection and interpretation of emotions [6], especially from appearances like facial expressions. With the increased popularity of video content in digital media, there is a growing interest in investigating techniques for identifying emotions in videos. Numerous emotional clues can be inferred from facial expressions, which range from delight and surprise to despair and rage. Combining computer vision and machine learning has greatly enhanced automated facial expression analysis, particularly in the field of video-based emotion identification [13]. However, there are difficulties in this endeavor. Accurate and reliable emotion recognition is hampered by the complexity of facial expressions, individual differences, occlusions, illumination variances, and the dynamic nature of emotions [11]. Moreover, temporal dependencies, nuanced facial dynamics, and the need for real-time analysis introduce additional complications to video content. With a focus on facial expression concepts, the difficulties in recognizing emotions in videos to improve interpretability and contextual understanding in emotion analysis [7], this research aims to investigate emotion detection in videos. The expressions on our faces, which are produced by a variety of facial muscles, are an essential means of conveying human emotions. Gaining an understanding of the principles underlying facial expressions is essential to correctly deciphering and classifying the emotions shown in video clips. Identifying emotions from video data poses a multitude of intricate difficulties. Significant challenges are created by the diversity of facial expressions people display, the coexistence of several emotions, and the fact that emotions change over time. Accurately identifying emotions in video content [9] is further complicated by environmental conditions, data noise, and occlusions. When opposed to still photos, videos present unique difficulties mostly because of timing dependencies. Emotion analysis over consecutive frames necessitates methods able to capture the dynamic nature of facial expressions. Resilient methods for feature extraction and classification across video frames are also necessary due to fluctuations in facial positions, motions, and occlusions. It may be possible to improve the analysis of emotions in films by incorporating natural language processing (NLP) tools [14]. NLP techniques provide additional cues for a more complex understanding of emotions and improve the overall precision of emotion detection systems by integrating contextual data, dialogues, or subtitles that accompany video content. In order to strengthen the interpretive power and robustness of emotion recognition models, this study will explore approaches and developments that address these issues in video-based emotion detection [10]. It will place particular emphasis on the combination of NLP techniques and the application of face expression concepts. Convolutional neural networks (CNNs) are one type of Deep Learning neural network design that is widely used in computer vision. "Computer vision" is the area of artificial intelligence that allows computers to interpret and process images and other visual data. In Machine Learning, Artificial Neural Networks exhibit remarkable performance. Neural networks are used in a number of datasets, including text, audio, and image datasets. Various neural network types are applied to various applications [12]. For example, convolution neural networks are utilized to classify images, whereas RNNs—more particularly, LSTMs—are used to predict word sequences.

A regular neural network has three main types of layers: Layers of Input: It is the layer where the data is fed. The total number of features in our data, or pixels in the case of an image, is same as the total number of neurons in this layer. Secret Layer: The input that was previously sent to the input layer is received by the concealed layer. Several hidden layers could exist, depending on our model and the volume of data.

Each hidden layer has a dissimilar number of neurons, but generally speaking, greater than the number of features. The output of each layer is determined by multiplying its predecessor's output by the learnable weights of that layer is a kind of neural network architecture made to manage sequential data by preserving a hidden state that records details about the sequence's earlier inputs. RNNs are very helpful in Natural Language Processing (NLP) applications involving language production and understanding, where word or character order is important [15] Here is an overview of the main ideas behind recurrent neural networks in natural language processing: Processing Sequential Data: RNNs work well when processing data in sets, such time series or sentences. They work on a single sequence element at a time, keeping a concealed state that contains data from earlier components. Hidden State: An RNN's hidden state acts as a memory for earlier inputs. Natural language processing (NLP)-based emotion identification in video is a significant and developing field of study with great potential across many areas. This innovative combination of computer vision, machine learning, and natural language processing aims to interpret the rich lexicon of human emotions as they are depicted in visual data. Within the framework of our work, we concentrate on the subtle interpretation of recorded facial expressions in order to understand the underlying affective states. There are several possible uses for emotion recognition in videos, including market research, psychology, human-computer interface, and healthcare. Comprehending human emotions is essential for developing more responsive and empathic technology systems, improving user experience, and allowing machines to interact with humans more deeply [8]. Research work goes beyond the conventional limitations of text-based analysis by expanding on the foundations of natural language processing (NLP) approaches to extract significant insights from visual data. We want to push the limits of accuracy and applicability in emotion identification from video information by using sophisticated machine learning models, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), k-Nearest Neighbors (KNN) and Recurrent Neural Networks (RNN). This work explores the methods used, the outcomes of the experiments, and the possible ramifications of our conclusions. Our goal is to further the field of emotion detection technology by investigating the combination of natural language processing and video analysis. This will enable the development of more advanced applications that will help close the gap between artificial intelligence and human emotions. The following sections include a review of the literature that identifies research gaps in the study, a methodology section that provides an overview of the proposed model and results, a discussion section that analyzes the results using figures and graphs, a conclusion, and future work that details the proposed work's conclusion and its future endeavor.

In the proposed work, the top 10 frames are taken into consideration every other second to analyze a person's emotion and forecast it using machine learning algorithms. To facilitate pre-processing and recognition, the top ten image frames are considered. The image frame is preprocessed to remove noise and enhance the quality of the image. In the end, the enhanced image frames are used to train and evaluate RNN, CNN, SVM, and KNN models in order to determine the image's emotion. Following are the contribution.

- The dataset consists of 32,298 different facial expression which includes seven categories (0=angry, 1=contempt, 2=disgust, 3=fear, 4=happy, 5=sadness, and 6=surprise).
- To identify the emotion of the person in the stored video using RNN and other machine learning algorithms.

II. Literature Survey

Wang, S. et al. [1] proposed a work that uses adversarial learning to systematically understand emotion distributions for classifying emotions in multimedia content is being done to solve this discrepancy. We use a discriminator and an emotion classifier in our technique. The discriminator distinguishes between expected and actual emotion labels, whereas the classifier predicts emotion labels by examining the multimedia content. The two components receive training at the same time, competing to improve their own performances. Sindhu, N. et al. [2] proposed research on a multimedia recommendation system powered by user emotions is presented in this research. The system chooses audio/video content automatically based on user emotions, eliminating the need for human browsing. The database is made up of ECG signals that were gathered from DECAF, with an emphasis on emotions that negatively affect mood, such as melancholy and rage. Raheel, A. et al. [3] aimed at enhancing the audience's emotional experience by utilizing rich multimedia content that stimulates their touch senses in addition to their visual and audio senses. The resulting p-values demonstrate substantial differences in valence and arousal levels between standard multimedia and TEM content, underscoring the greater emotional impact of TEM clips. Twelve temporal domain variables are taken out from the preprocessed EEG input for emotion recognition, and four human emotions—happy, furious, sad, and disgusted—are classified using a support vector machine. Zhang, X. et al. [4] proposed research on augment visual and aural perceptions with tactile multimedia content in order to increase the audience's emotional engagement. Using a t-test on valence and arousal levels highlights important differences between TEM and traditional multimedia, highlighting the higher emotional resonance of TEM content. In the field of emotion recognition, the preprocessed Four human emotions—happy, angry, sad, and calm—are classified by a support vector machine using the twelve temporal domain variables that an EEG signal produces. The remarkable results of 43.90% and 63.41% accuracy against TEM clips and standard multimedia, respectively, demonstrate the improved efficacy of EEG-based emotion recognition, particularly in stimulating the touch sense. Bhattacharya, S. et al. [5] focused on the research emotion detection in online social networks (OSNs) has the potential to enhance several applications, including targeted advertisement services and recommendation systems. Emotion analysis has historically focused mostly on identifying single emotion labels or predicting sentence-level polarity, ignoring the potential of many coexisting emotions from the viewpoint of users. In this work, we refer to the topic as a multilabel learning challenge and tackle multi-emotion recognition in

OSNs from a user-level perspective. First, we use an annotated Twitter dataset to investigate relationships between emotion labels, social ties, and temporal patterns. Chen, L. et al. [15] gives a summary of current work in the rapidly developing topic of automatic group emotion identification is given in this article. Research has used a variety of datasets, modalities (video, pictures, social media posts, audio), and approaches to investigate emotion analysis in crowds or groups. When possible, we want to provide code access and implementation details. Subsequent research endeavors will focus on developing real-world-applicable systems, accommodating varying group sizes, affective subsets, and affective evolution, enhancing resilience, and employing less biased datasets. Abdu, S. A. et al. [16] centered on a novel approach that uses Bag-of-Audio-Words (BoAW) to extract features from conversational audio data. It offers an advanced Recurrent Neural Network (RNN)-based emotion recognition model that is great at forecasting the present emotions while also capturing the emotional states of the participants and the context of the conversation. The effectiveness of the strategy is demonstrated by experiments on two benchmark datasets and realistic real-time assessments. This strategy outperforms the current state-of-the-art models with weighted accuracy of 60.87% and unweighted accuracy of 60.97% for six basic emotions on the IEMOCAP dataset. Chen, L. et al. [17] reports on our preliminary investigation of sophisticated multimodal emotion detection techniques used to evaluate interviewee performance in emotionally charged situations. Although the results point to the potential of FACET in emotion recognition, there don't seem to be many advantages to using SER. Fernandes, R et. al. [18] focused on video captures the human face in action and provides further insights into human emotions, we have utilized emotion recognition to analyze human emotions in this work. Deep learning algorithms are applied in this article to identify human emotions from archived video footage. In an effort to predict the many emotions shown in a stored video—namely, anger, surprise, happiness, and neutrality—we have looked at Convolutional Neural Networks (CNNs). Wei, J. et al. [19] proposed an affective saliency estimation-based key frames extraction approach. Key frames are extracted from videos by calculating their emotive saliency value in order to prevent emotion-neutral frames from affecting the recognition outcome. To extract useful deep visual information, pretrained models and conventional models are employed. Random Forests (RF), Support vector machines (SVM), and deep model Convolutional Neural Networks (CNN) are used to recognize emotions. Washington, P. et al. [20] suggested a secure web-based platform that transforms manual labeling work into an activity. We gathered and tagged 2,155 films. However, 79.1% balanced accuracy and 78.0% F1-score were obtained for the CAFE Subset A, which had at least 60% human agreement on emotion labels. Siam, A. I. et. al. [21] Based on the real-time deep learning-based MediaPipe face mesh technology, the main concept is to produce key locations. To further encode the generated key points, a number of well-designed mesh generator and angular encoding modules are employed. Moreover, feature decomposition employs Principal Component Analysis (PCA). Gunawan, T. S. et al. [22] focused on the process of identifying emotions from films using deep learning algorithms. This paper presents the methodology and explanation of the recognition process. We also look at some of the video-based datasets that are utilized in numerous academic publications. The performance metrics of the various emotion recognition models are shown with the results. Results from an experiment on depression detection using Google 97% accuracy on the training set and 57.4% accuracy on the testing set were demonstrated by Colab's fer2013 dataset. Jaiswal, A. et al. [23] The development of a system that recognizes emotions in facial expressions using artificial intelligence (AI) is underway. The three primary processes in the emotion detection process face identification, feature extraction, and emotion classification are covered. To identify emotions from photos, this study presented a deep learning architecture based on convolutional neural networks (CNNs). Two datasets are utilized to examine the performance of the advised method: the Japanese Female Facial Emotion (JAFFE) and the Facial Emotion Recognition Challenge (FERC-2013). For the FERC-2013 and JAFFE datasets, the accuracy percentages yielded by the suggested model were 70.14 and 98.65, respectively. Mukhopadhyay, M. et al. [24] four intricate feelings were chosen from an actual poll. Basic human emotions are grouped together to form complex emotions, which are frequently felt by a group of students throughout a lesson. Rather of using discrete images, we thought of using a predetermined collection of continuous image frames to accurately convey these related feelings. In order to categorize the fundamental emotions and determine the learners' mental states, we constructed a CNN model. Both a mathematical verification and a learner survey are used to confirm the results. The findings indicate that the accuracy rates for classifying emotions and identifying states of mind are 65% and 62%, respectively. Mehendale, N. et al. [25] The FERC is based on research on two-part convolutional neural networks (CNNs). In the first section, the image's backdrop is removed, and in the second, the focus is on facial feature vector extraction. The FERC model uses the expressional vector (EV) to distinguish between five different types of regular facial expression. From the 10,000 images (154 individuals) that were stored in the database, supervisory data was obtained. It is possible to illuminate the emotion with 96% accuracy when the EV of length 24 values is used. While the two-level CNN functions in series, the last layer of the perceptron alters the weights and exponent values for each iteration. FERC improves accuracy by deviating from commonly used single-level CNN techniques. Additionally, a unique background cleaning process used prior to EV creation prevents. Yadav et al. in study [26], talks about the survey article attempts to provide reviews of the most recent machine learning architectures, the applications of the system, the use of algorithms, and speech and vision processes. The technology of today presents enormous research opportunities.

Algorithms and architectures in machine learning can also be intelligently applied to generate new ideas and intelligently replicate speech and vision systems. The degree of commercialization and personal computing is at an all-time high. Huge amounts of sensor data and cloud computing can be used for machine learning and training. Even more advanced technologies can be found in mobile and embedded systems. Abdullah et al. [27] A vital component of human communication is facial expression. Thus, accurate facial

expression classification in picture and video data has become a major research goal for the software development community. In this work, we offer a method for classifying videos by capturing both the temporal and spatial aspects of a video sequence using Recurrent Neural Networks (RNN) in addition to Convolution Neural Networks (CNN). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is used to test the method. Assuming that visual analysis was the sole technique available to gather data on this dataset, the proposed approach yields the initial benchmark of 61% test accuracy. Awais et al. [28], the suggested system provides emotional recognition and real-time communication, allowing for remote learning support and health monitoring during pandemics. The study's obtained results show great promise. The proposed IoT protocols, TS-MAC and R-MAC, achieve an ultralow latency of 1 ms. Furthermore, reliability is improved by R-MAC in comparison to state-of-the-art. Moreover, the proposed deep learning method achieves remarkable performance, with an f-score of 95%. The outcomes in the domains of AI and communications are consistent with the interdependency requirements of deep learning and Internet of Things frameworks. This ensures that the suggested task is suitable for usage in healthcare, student engagement, emotion support, remote learning, and general wellness. Zahara et al. [29] This study recommends developing a system that extracts characteristics from facial expressions and utilizes the Convolution Neural Network (CNN) technique to classify them in real-time. It does this by utilizing the OpenCV library's TensorFlow and Keras. The study design used with the Raspberry Pi consists of three main processes: face detection, face feature extraction, and facial emotion classification. 65.97% (sixty-five point ninety-seven percent) of the facial expressions in the study utilizing the Facial Emotion Recognition (FER-2013) Convolutional Neural Network (CNN) technique were predicted. Mohan et. al. [30] proposed FER-net, a convolution neural network that efficiently differentiates FEs by using the softmax classifier. We implement our method, FER-net, and assess it on five benchmarking datasets: FER2013, Extended Cohn-Kanade, Karolinska Directed Emotional Faces, Real-world Affective Faces, and Japanese Female Facial Expressions. We also use twenty-one other state-of-the-art methodologies. The seven FEs—neutral, anger, disgust, fear, pleasure, sorrow, and surprise—are examined in this essay. The average accuracy ratings for these datasets are 81.68%, 96.7%, 97.8%, 82.5%, and 78.9%, in that order. The obtained findings demonstrate the superiority of FER-net over twenty-one state-of-the-art methods. Unlike prior studies that concentrated primarily on real-time streaming videos with audio, the current work aims to discern people's moods from recorded videos. The proposed work aims to recognize people's emotions from recorded films, which distinguishes us from other studies that just focused on videos streamed in real time with a limited range of emotion classifications. Unlike previous work that employed webcams and a convolutional neural network, our method explores a wide range of machine learning techniques to improve the accuracy of emotion identification.

III. METHODOLOGY

Fig. 1 depicts the block diagram of proposed model. This section covers the methodology used for the emotion detection such as Data collection, Image Preprocessing, Feature Extraction, Classification using machine learning algorithms and result analysis.

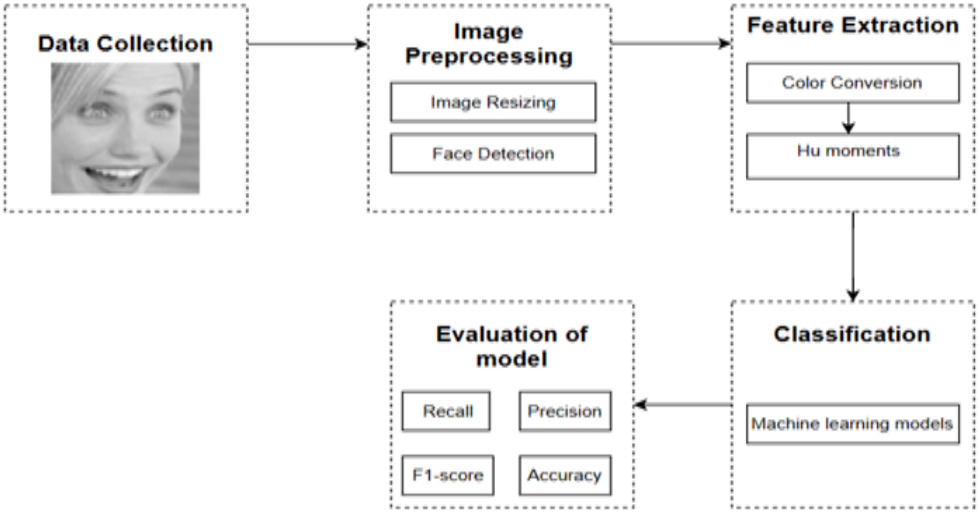

Fig. 1. Block diagram of proposed model.

A. Data Collections

The data set was gathered from https://www.kaggle.com/datasets/msambare/fer2013 on Kaggle. The videos are in the mp4 format and have been resized to 320*240 pixels. Each video has a duration of twenty seconds. The top ten frames will be taken into consideration after the model has received video input per second.

B. Image Preprocessing

Using face detection techniques, such Haar cascades, to locate and extract face regions from an image is a basic first step. These techniques are essential for precisely identifying faces in a range of positions and angles. Aligning the identified faces to a consistent orientation is a crucial preprocessing step once faces are discovered. Because it minimizes variation in

face positions and increases the general dependability of ensuing analysis, this alignment procedure is essential for guaranteeing data consistency. Aligning the faces enables more reliable and precise results by allowing the later phases of facial recognition and feature extraction to function on uniform facial structures.

C. Feature Extraction

From the previously processed images, extract pertinent features that effectively capture facial expressions. Facial landmarks, texture features, or other features can be used as these, or deep learning-based feature representations using convolutional neural networks (CNNs) or other machine learning techniques used in this study. The main process in the feature extraction is color conversion where the colored image is being converted into the gray image and then feature is extracted by using Hu moments feature. Hu moments feature are used where the shape of the face is being identified.

Algorithm 1: Feature Extraction algorithm
Read the input image from the specified image_path Resize the image to 320x240 pixels. Load the pre-trained Haar cascade classifier for face detection. Convert the resized image to grayscale. Detect faces in the grayscale image using the Haar cascade classifier. Initialize an empty list to store extracted features For each detected face: a. Crop the face region from the grayscale image. b. Calculate moments for the cropped face region. c. Calculate Hu moments from the moments. d. Add the extracted features to the list. Output the extracted features for each detected face

D. Classification Using Machine Learning Algorithms

To train the model, 16861 facial expression photos were used in the dataset. The labels that corresponded to the images were used to obtain them. Testing and training sets were created from the dataset. To reduce the possibility of overfitting, the model, which was built using the Keras sequential technique, uses dropout to randomly deactivate specific neurons. The dataset, which was divided into seven classes: anger, contempt, disgust, fear, happiness, sadness, and surprise was used to train the model. The provided dataset was employed to train a convolutional neural network. To better understand multi-class categorization, we looked at seven different classes: Happiness, anger, surprise, sadness, contempt, disgust, and fear.

E. Recurrent Neural Network

Among the Neural Network types that tackle the problem of predicting the next word in a sequence, the Recurrent Neural Network (RNN) stands out. In contrast to conventional neural networks, which function independently of inputs and outputs, Because of their architecture, RNNs can remember words that have come before them, which is a vital feature for tasks that require sequential prediction. The most important component that makes this memory function possible is the Hidden Layer in RNNs. In some applications, this layer resolves the need on previous context by maintaining crucial information about a sequence and differentiating RNNs. The state, sometimes known as the Memory State, keeps data from the prior input in the network. It reduces parameter complexity compared to other neural networks by using the same parameters for every input or hidden layer, which allows it to do the same task for every input. Fig. 2 depicts the block diagram of recurrent neural network. Table I and II provides the details of hyperparameter values of CNN and RNN algorithms.

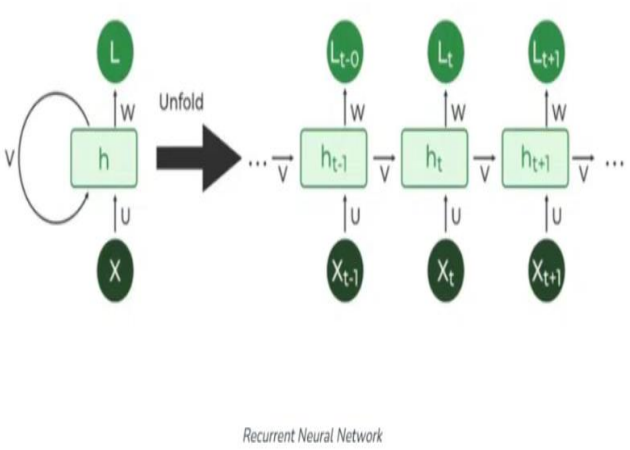

Fig. 2. Block diagram of recurrent neural network.

TABLE I. HYPERPARAMETERS OF RECURRENT NEURAL NETWORK

Hyperparameters	Values
hidden_size	128
num_layers	2
learning_rate	0.001
dropout	0.5
batch_size	64
seq_length	20
optimizer	‘Adam’
activation	'tanh'

TABLE II. HYPER PARAMETERS OF CONVOLUTION NEURAL NETWORK

Hyper parameters	Values
Learning Rate	0.1
Number of Epochs	100
Batch Size	64
Filter Size	3x3
Pooling Type	Max
optimizer	'Adam
activation	'tanh’

IV. RESULT AND DISCUSSION

Fig. 3 illustrates the Loss analysis model plotted against epochs, demonstrating favorable outcomes on validation data. Epochs of 100 is considered for the analysis of loss model. Fig. 4 illustrates the Accuracy analysis model plotted against epochs, demonstrating favorable outcomes on validation data.

Epochs of 100 is considered for the analysis of the Accuracy model. ROC graph is plotted by considering the 7 classes. A graph illustrating precision versus recall is generated for the seven classes. The four classes are represented, with class 0 corresponding to anger, class 1 to contempt, class 2 to disgust, class 3 to fear, class 4 to happy, class 5 to sadness, and class 6 to surprise. Curves that deviate from the baseline indicate higher potential levels, as shown in Fig. 5.

Fig. 7 Performance Evaluation Matrix The confusion matrix is graphed to assess the effectiveness of N classes, forming an NxN matrix. In this analysis, we consider 7 classes, denoted as 0=angry, Sure, here is a rearrangement of the words: 1=happy, 2=disgust, 3=fear, 4=contempt, 5=surprise, 6=sadness. The matrix serves to contrast predicted values with actual values, as illustrated in Fig. 6. Fig. 6 compares several models, including the 95% accurate Recurrent Neural Network (RNN), the 92% accurate onvolution Neural Network (CNN), the 90% accurate Support Vector Machine, and the 87% accurate K Nearest Neighbor. Fig. 8 compares the performances of several models using characteristics like F1-score, accuracy, precision, and recall. Table III shows the performance metrics of different algorithms.

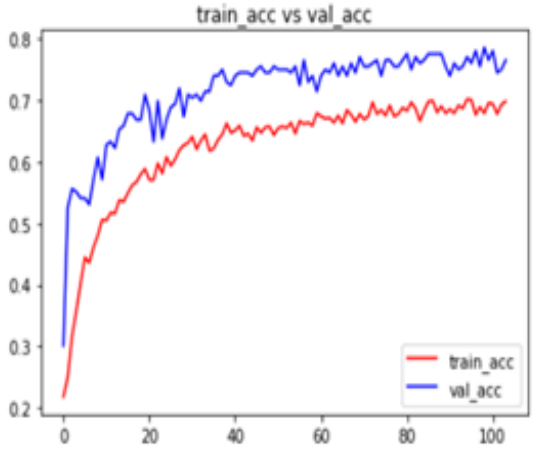

Fig. 3. Loss analysis model.

Fig. 4. Accuracy analysis model.

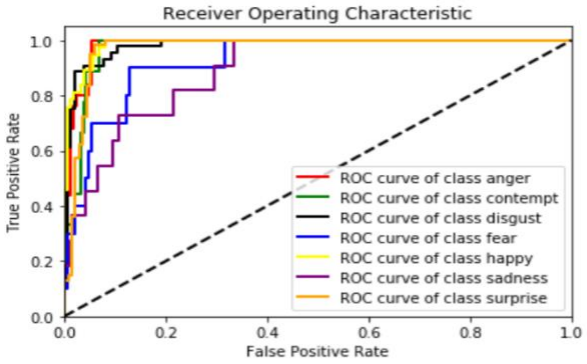

Fig. 5. ROC curve graph.

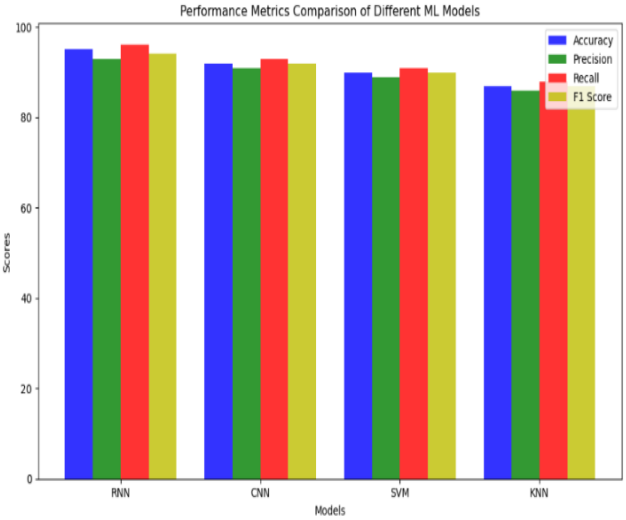

Fig. 6. Performance metrics of different models.

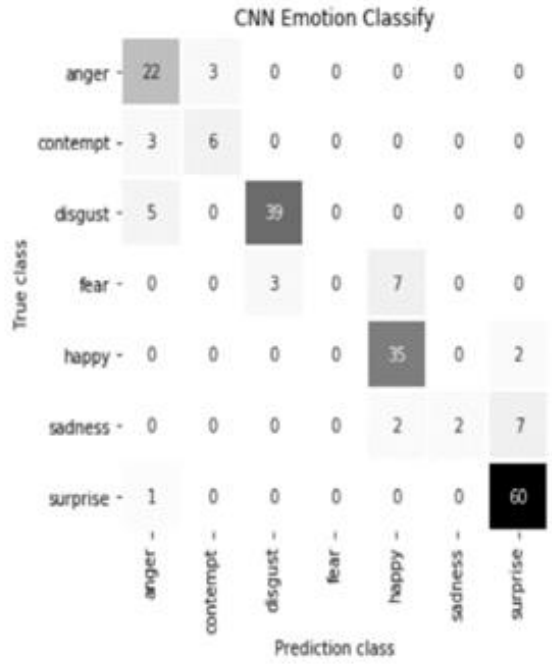

Fig. 7. Confusion matrix.

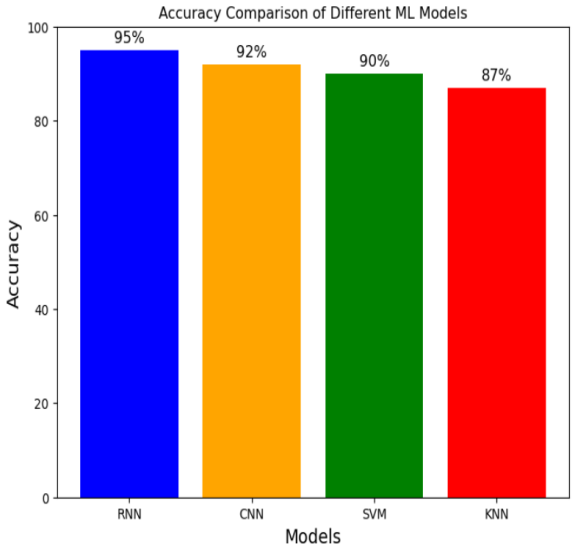

Fig. 8. Model accuracy comparisons.

TABLE III. PERFORMANCE METRICS OF DIFFERENT ALGORITHMS

Algorithm	Precision (%)	Accuracy (%)	Recall (%)	F1-Score (%)
RNN	93	95	96	94
CNN	91	92	93	92
SVM	89	90	91	90
KNN	86	87	88	87

V. CONCLUSION AND FUTURE WORK

The results of the studies emphasize the significant potential that Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and RNN possess. These models show impressive skills that show how well they can decode and interpret the complex web of facial expressions. The decision to employ the RNN model is guided by its intrinsic sequential nature, aligning seamlessly with the temporal dynamics present in video data. Impressively, the RNN model attains a commendable 95% accuracy rate, attesting to its proficiency in discerning and categorizing emotional nuances portrayed through facial expressions. This high accuracy rate lays the groundwork for potential expansions of the project, such as predicting varying levels of emotions like playfulness and delight. Looking ahead, the project could evolve towards predicting emotional states, offering a more nuanced understanding of human affect.

A strategic step towards refining the model involves its application across diverse datasets, ensuring that it can adapt and perform optimally across a spectrum of scenarios. This iterative process not only enhances the model's accuracy but also positions it as a robust tool for real-world applications in emotion recognition technology.

REFERENCES

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

[2] Wang, S., Peng, G., Zheng, Z., & Xu, Z. (2019). Capturing emotion distribution for multimedia emotion tagging. IEEE Transactions on Affective Computing, 12(4), 821-831.

[3] Sindhu, N., Jerritta, S., & Anjali, R. (2021, February). Emotion driven mood enhancing multimedia recommendation system using physiological signal. In IOP Conference Series: Materials Science and Engineering (Vol. 1070, No. 1, p. 012070). IOP Publishing.

[4] Raheel, A., Anwar, S. M., & Majid, M. (2019). Emotion recognition in response to traditional and tactile enhanced multimedia using electroencephalography. Multimedia tools and applications, 78(10), 13971-13985.

[5] Zhang, X., Li, W., Ying, H., Li, F., Tang, S., & Lu, S. (2020). Emotion detection in online social networks: a multilabel learning approach. IEEE Internet of Things Journal, 7(9), 8133-8143.

[6] Bhattacharya, S., Borah, S., Mishra, B. K., & Mondal, A. (2022). Emotion detection from multilingual audio using deep analysis. Multimedia Tools and Applications, 81(28), 41309-41338.

[7] Raheel, A., Majid, M., Alnowami, M., & Anwar, S. M. (2020). Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia. Sensors, 20(14), 4037.

[8] Li, M., Xie, L., Lv, Z., Li, J., & Wang, Z. (2020). Multistep deep system for multimodal emotion detection with invalid data in the internet of things. IEEE Access, 8, 187208-187221.

[9] Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN.

[10] Veltmeijer, E. A., Gerritsen, C., & Hindriks, K. V. (2021). Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing, 14(1), 89-107.

[11] Rana, A., & Jha, S. (2022). Emotion based hate speech detection using multimodal learning. arXiv preprint arXiv:2202.06218.

[12] Huang, X., Ren, M., Han, Q., Shi, X., Nie, J., Nie, W., & Liu, A. A. (2021). Emotion detection for conversations based on reinforcement learning framework. IEEE MultiMedia, 28(2), 76-85.

[13] Qi, F., Yang, X., & Xu, C. (2020). Emotion knowledge driven video highlight detection. IEEE Transactions on Multimedia, 23, 3999-4013.

[14] Veltmeijer, E. A., Gerritsen, C., & Hindriks, K. V. (2021). Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing, 14(1), 89-107.

[15] Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., & Nanayakkara, V. (2022). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. Multimedia Tools and Applications, 81(24), 35173-35194.

[16] Chen, L., Yoon, S. Y., Leong, C. W., Martin, M., & Ma, M. (2014, November). An initial analysis of structured video interviews by using multimodal emotion detection. In Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems (pp. 1-6).

[17] Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. Information Fusion, 76, 204-226.

[18] Chen, L., Yoon, S. Y., Leong, C. W., Martin, M., & Ma, M. (2014, November). An initial analysis of structured video interviews by using multimodal emotion detection. In Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems (pp. 1-6).

[19] Fernandes, R., & Rodrigues, A. P. (2022, December). Emotion Detection in Multimedia Data Using Convolution Neural Network. In 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE) (pp. 157-161). IEEE.

[20] Wei, J., Yang, X., & Dong, Y. (2021). User-generated video emotion recognition based on key frames. Multimedia Tools and Applications, 80, 14343-14361.

[21] Washington, P., Kalantarian, H., Kent, J., Husic, A., Kline, A., Leblanc, E., ... & Wall, D. P. (2020). Training an emotion detection

classifier using frames from a mobile therapeutic game for children with developmental disorders. arXiv preprint arXiv:2012.08678.

[22] Siam, A. I., Soliman, N. F., Algarni, A. D., El-Samie, A., Fathi, E., & Sedik, A. (2022). Deploying machine learning techniques for human emotion detection. Computational intelligence and neuroscience, 2022.

[23] Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. (2020). Development of video-based emotion recognition using deep learning with Google Colab. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(5), 2463-2471.

[24] Jaiswal, A., Raju, A. K., & Deb, S. (2020, June). Facial emotion detection using deep learning. In 2020 international conference for emerging technology (INCET) (pp. 1-5). IEEE.

[25] Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P. K. D., Dasgupta, N., & Choudhury, P. (2020, February). Facial emotion detection to assess Learner's State of mind in an online learning system. In Proceedings of the 2020 5th international conference on intelligent information technology (pp. 107-115).

[26] Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2(3), 446.

[27] Yadav, S. P., Zaidi, S., Mishra, A., & Yadav, V. (2022). Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Archives of Computational Methods in Engineering, 29(3), 1753-1770.

[28] Abdullah, M., Ahmad, M., & Han, D. (2020, January). Facial expression recognition in videos: An CNN-LSTM based model for video classification. In 2020 International Conference on Electronics, Information, and Communication (ICEIC) (pp. 1-3). IEEE.

[29] Awais, M., Raza, M., Singh, N., Bashir, K., Manzoor, U., Islam, S. U., & Rodrigues, J. J. (2020). LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19. IEEE Internet of Things Journal, 8(23), 16863-16871.

[30] Zahara, L., Musa, P., Wibowo, E. P., Karim, I., & Musa, S. B. (2020, November). The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi. In 2020 Fifth international conference on informatics and computing (ICIC) (pp. 1-9). IEEE.