WCRG Global Research
WORLD COMET RESEARCH GROUP

International Journal of Artificial Intelligence & Digital Transformation
Vol. 8, No. 1, 2025
Doi: 10.67228/30713315/IJAIDT-2025PI3V1Z
PP. 01-12

Original Article

Deep Learning Approaches for Real-Time Emotion Recognition

Dr. Venkatesh Iyer

Professor, SRM Institute of Science and Technology, India.

Received: 12-02-2024 Revised: 12-25-2024 Accepted: 01-02-2025 Published: 01-04-2025

ABSTRACT

Real-time emotion recognition has become an important area in affective computing due to advances in deep learning and the need for intelligent human-computer interaction. This paper reviews pre-2019 deep learning approaches for emotion recognition using modalities like facial expressions, speech, and multimodal data. Traditional methods using handcrafted features (LBP, HOG, MFCC) had limited performance, while deep learning models such as CNNs, RNNs, LSTMs, and DBNs improved accuracy through automatic feature extraction and temporal learning. CNNs performed well in facial recognition tasks, while RNNs and LSTMs were effective for speech analysis. Multimodal systems showed higher accuracy but faced challenges like synchronization and complexity. Despite improvements, real-time deployment remains difficult due to high computational demands, cultural variability, and environmental noise. Overall, deep learning has significantly advanced emotion recognition, but challenges like efficiency, generalization, and interpretability still need to be addressed.

KEYWORDS

Emotion Recognition, Deep Learning, CNN, RNN, LSTM, Affective Computing, Real-Time Systems, Facial Expression Analysis, Speech Processing, Multimodal Learning.

WCRG
Global Research
ISSN: 3071-3315
PP: 01-12.

1. INTRODUCTION

1.1. Background

The need to understand human emotional states and respond to them has made emotion recognition an interesting field of study because it could potentially enable machines to understand and respond to human emotions, thereby greatly improving the interaction between humans and computers. Virtual assistants, healthcare monitoring systems, and adaptive learning platforms are just some of the applications that require the accuracy of emotion detection. The systems used in the past were based on rule-based algorithms and hand-crafted characteristics such as Local Binary Patterns and Mel-Frequency Cepstral Coefficients, which demanded domain expertise, and did not always work well on different data sets and real world conditions. These techniques had weaknesses in dealing with differences in light, speech patterns, and personal expressions. With the development of deep learning, especially deep learning models like Convolutional neural networks, this area has been changed as they allow automatic feature extraction and hierarchical learning on raw data. This change has made emotion recognition systems more transparent, scalable, and adaptable to better suit real world applications.

1.2. Importance of Real-Time Emotion Recognition

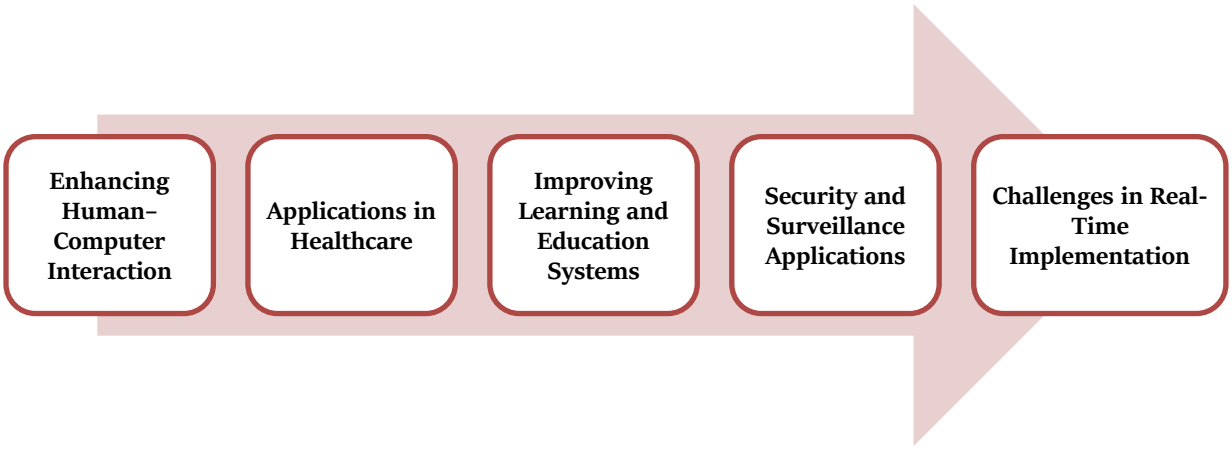

Fig 1 - Importance of Real-Time Emotion Recognition

1.2.1. Enhancing Human–Computer Interaction

Emotion recognition in real-time leads to a large enhancement in the quality of human-computer interaction because it allows the system to react immediately to the emotional state of users. Rather than giving a generic answer, intelligent systems are capable of changing their mode of operation depending on the emotions that have been identified like frustration, happiness or confusion. This results in more natural, interactive and personalized interactions, particularly in applications such as virtual assistants, chatbots, and interactive systems that run on Artificial Intelligence.

1.2.2. Applications in Healthcare

Real-time emotion recognition is an important aspect in the healthcare field where it is used to monitor the mental and emotional well-being of patients. A system can detect facial expressions or

ISSN: 3071-3315
PP: 01-12.

speech patterns that may indicate stress, anxiety, or depression and notify the caregiver or medical professional. It is effective especially in remote healthcare and geriatric care systems where constant surveillance could save the lives of patients and offer prompt interventions.

1.2.3. Improving Learning and Education Systems

Emotion detection in real-time improves e-learning systems to adjust the presentation of the material in accordance with the emotional reactions of students. As an example, the system can change the level of difficulty or elaborate further or offer other ways to learn in case a student seems confused or inattentive. This makes the learning process more efficient and personalized, making the students more engaged and retaining more knowledge.

1.2.4. Security and Surveillance Applications

Security and surveillance can also use emotion recognition to detect suspicious behavior or an abnormal behavior. These systems can help to avert possible threats or criminal acts by detecting emotional signals, like fear, anger, or nervousness, in real time. This is not a foolproof feature, but this feature provides an extra level of intelligence to the current surveillance technologies.

1.2.5. Challenges in Real-Time Implementation

Irrespective of its significance, real-time emotion recognition encounters issues like computational complexity, latency, and environmental variability. The algorithms and optimized hardware are needed to process large amounts of data within a short period of time, with high accuracy. Also, lighting variations, background noise, individual expression style can influence system performance, and it is necessary to create powerful and flexible models.

1.3. Challenges in Emotion Recognition

Recognition of emotions is a complicated process, as it is necessary to interpret human expressions, which can be very subtle and even ambiguous, and high accuracy and reliability are hard to achieve. Among the key problems is the differences in the emotional expressions of different people. Emotions are different among people depending on the culture, personality as well as the situation, and this complicates the generalization of models across different populations. Moreover, not all the emotions are expressed in a clear and exaggerated way; sometimes it is more complicated as some expressions are subtle, or mixed feelings can be detected. The other significant issue is the environmental factors like low light, covering (e.g. glasses, masks) and noises of audio signals which can significantly impair the quality of input data and influence the work of models. Technically, it is hard to capture spatial and temporal aspects of emotions. Models such as Convolutional Neural Networks are good at extracting spatial features of images; however, they can be poor at capturing temporal dynamics unless used with sequence models. In contrast, other models like the Long Short-Term Memory are set up to handle sequential data, but are computationally expensive and need large quantities of labeled data. The lack of data and its disproportions are also a problem, because not all emotional classes can be represented, and it can result in biased forecasts. In addition, real-time implementation also presents limitations regarding the speed of processing and availability of

ISSN: 3071-3315
PP: 01-12.

resources, particularly the embedded system or a mobile system. Ethical issues, such as the problem of privacy and the possibility of exploiting the emotion recognition technologies, further complicate the matter. On the whole, these issues indicate that more powerful, effective, and ethically accountable methods of emotion recognition research are required.

2. LITERATURE SURVEY

2.1. Early Feature-Based Methods

Until the emergence of deep learning, emotion recognition systems mainly relied on manually designed features derived out of images, audio, or video. Local Binary Patterns (LBP) techniques were employed to reproduce fine-grained texture features such as wrinkles and micro-expressions on the face, but were very vulnerable to noise and light changes. Likewise, Histogram of Oriented Gradients (HOG) specialized in edge and shape data, which performed well with stationary images but failed to distinguish between changes in expression over time. Mel-Frequency Cepstral Coefficients (MFCC) were popular in the audio field to characterize speech signals but they did not have contextual or semantic knowledge of emotions over time.

2.2. Deep Learning Approaches

2.2.1. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) proved to be the breakthrough in facial emotion recognition as it has the capability of auto-learns hierarchical feature representations on raw images. They are good at robustly describing spatial patterns like facial features, expressions and textures without being translation-dependent, i.e. they are able to find features irrespective of their location in the image. This not only removed the manual feature engineering but also greatly enhanced accuracy in image-based emotion detection tasks.

2.2.2. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are structured to handle a sequence of data by storing a hidden state that stores information at previous time steps. RNNs in emotion recognition RNNs come in handy in situations where the emotion being studied is conveyed through speech or video sequence where it fluctuates with time. Nevertheless, the classic RNNs are characterized by such problems as the disappearance of the gradient, which complicates the ability to acquire long-term relationships.

2.2.3. Long Short-Term Memory (LSTM)

To overcome the vanishing gradient problem, Long Short-Term Memory (LSTM) networks were proposed as an enhancement of regular RNNs. In their operation, they have specialized gating processes, namely, input and output and forget gates, that regulate the passage of information so that the network can store significant emotional content across extended sequences. This renders LSTMs very useful in emotion recognition tasks of speech and video.

ISSN: 3071-3315
PP: 01-12.

2.2.4. Deep Belief Networks (DBN)

One of the earlier deep learning models that are applied to emotion recognition is Deep Belief Networks (DBNs). They are composed of layers of restricted Boltzmann machines, stacked together to learn to encode complex patterns in data by unsupervised pretraining. Despite the promising results with DBNs, they were later surpassed by better architectures such as CNNs and LSTMs.

2.3. Multimodal Emotion Recognition

Multimodal emotion recognition combines knowledge of more than one data source, to enhance reliability and precision. This method usually involves a combination of visual information (facial expressions), auditory information (Speech tone and pitch), and physiological information (heart rate or skin conductance). Complementary information across modalities can enable systems to more effectively capture the complexity of human emotions and process ambiguous or noisy inputs, which single-modality systems cannot.

3. METHODOLOGY

3.1. System Architecture

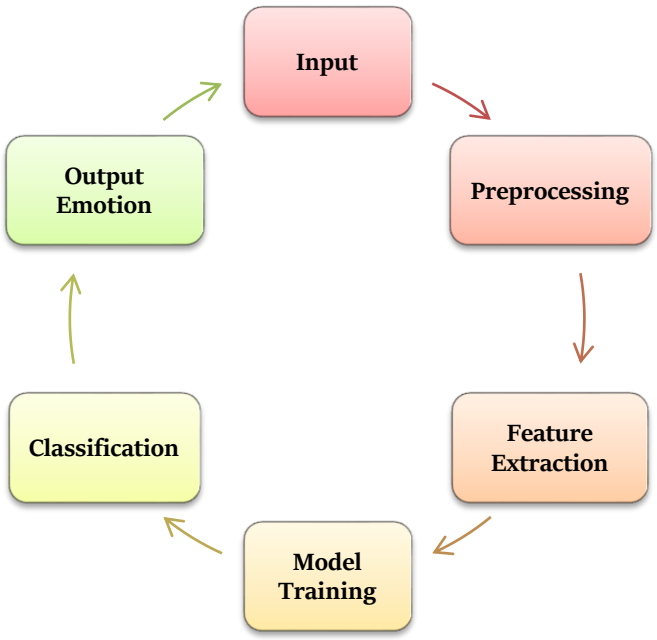

Fig 2 - System Architecture

3.1.1. Input

The input stage comprises the gathering of the raw data in various forms including images, video streams or audio records. In emotion recognition systems, this can be in the form of facial images of a camera, speech data of a microphone or a combination of the two in multimodal data. The nature and quality of the input data is very important in the overall performance of the system because dirty or incomplete data can largely impact the downstream processing.

ISSN: 3071-3315
PP: 01-12.

3.1.2. Preprocessing

The raw input data is preprocessed to allow the input data to be analyzed by filtering out noise and standardizing the format. In the case of images and videos, it can include face detection, resizing, normalization, and illumination correction. In the case of audio data, noise filtering, silence removal, and signal normalization are part of preprocessing. This will be done so that the irrelevant variation is kept to the minimal and the model concentrates on important emotional cues.

3.1.3. Feature Extraction

During this step, critical features are learned on the transformed data to reflect emotional data well. The classical methods made use of manual features and the newest systems make use of automated feature learning in models such as the Convolutional Neural Networks in visual data or the Mel-Frequency Cepstral Coefficients in speech signals. The aim is to reduce raw data to a concise and informative format that indicates emotion-related patterns.

3.1.4. Model Training

The extracted features are fed to machine learning or deep learning algorithms in model training to learn patterns that are related to various emotions. Sequential data are frequently trained with Techniques like Recurrent Neural Networks and Long Short-Term Memory whereas Image-based tasks rely on CNNs. In this stage, the model also tweaks its internal parameters, with labeled data to help reduce the error in prediction and enhance accuracy.

3.1.5. Classification

The classification step is the process in which the trained model is used to make the prediction of the emotional state using the learned features. The system identifies the input data as falling into predetermined categories of emotions, like happiness, sadness, anger, or surprise. The activation functions and decision layers (e.g., softmax) are usually used in this step to compute probability scores of all classes of emotions and choose the most probable one.

3.1.6. Output Emotion

The last step gives the identified feeling as the system output. Real-time applications can show this as a label, probability distribution or a visual indicator. In real-world systems, the result can also be decision-based, like the user interface, improvement of human-computer interaction, or to assist applications in health, education, and entertainment.

3.2. Data Preprocessing

Preprocessing of data is an important process of the emotion recognition systems because it makes sure that raw data received is processed into clean, consistent and analysis ready data. Image normalization is one of the main operations of preprocessing whereby the values of pixel intensity are adjusted to a standard scale. This aids in minimizing the differences in lighting, camera quality or contrast differences among images. Normalization can be performed using min-max scaling or z-score normalization to make all input images have an even distribution of pixel values, enabling

ISSN: 3071-3315
PP: 01-12.

models such as Convolutional Neural Networks to learn more effectively without being skewed by variations in intensity. Noise removal is another vital element that is aimed at removing unwanted distortions that are introduced in the data. In pictures, noise can be due to low light, sensor defects, or compression artifacts, whereas in audio data, noise can be due to the background sounds or interference during the recording process. Smoothing algorithms or median filtering or Gaussian filtering are some techniques typically applied to clean the data. The elimination of noise is significant because it provides an opportunity to highlight the significance of key aspects, including facial expression or speech pattern, and, therefore, increases the effectiveness of the system in recognizing emotions. Another preprocessing step is feature scaling which makes all features extracted have equal contribution to the learning process. Given that various features could be in various ranges and units, scaling techniques such as standardization or normalization are used to bring them to a similar range. This avoids dominance of features with higher numerical values in the learning process to enhance the learning rate of machine learning models. On the whole, preprocessing is a crucial step in improving the quality of data and minimizing variability, as well as the speed and accuracy of emotion recognition systems.

3.3. Model Design

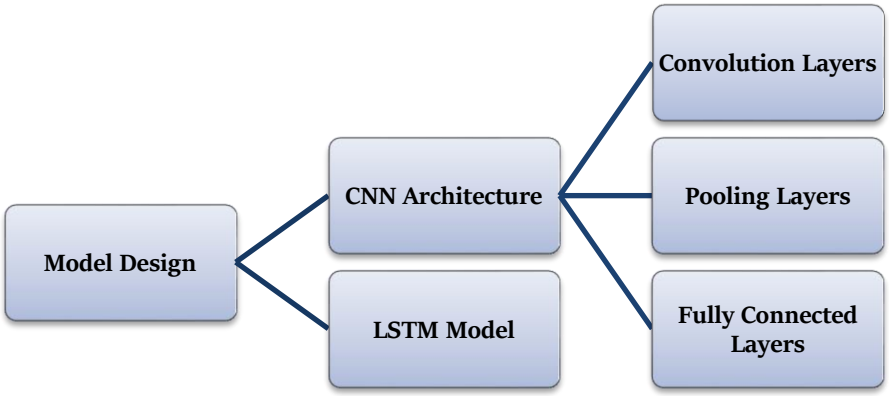

Fig 3 - Model Design

3.3.1. CNN Architecture

- Convolution Layers: Convolution layers are the heart of a Convolutional Neural Networks (CNN) and are the ones that perform the identification of significant spatial properties within the input images. These layers use several learnable filters (kernels) that move across the image and identify patterns within the image like edges, textures, and facial features. The deeper the network is, the more complex and abstract representations the convolution layers capture, e.g. expressions that are related to emotions and hence, most effective in tasks of facial emotion recognition.
- Pooling Layers: Reducing the spatial dimensions of feature maps produced by convolution layers in pooling layers helps to reduce computational complexity and eliminate overfitting. The most significant information in a region of the feature map is summarized by common techniques, like max pooling and average pooling. The process not only aids in the

ISSN: 3071-3315
PP: 01-12.

preservation of the dominant features and the elimination of less relevant features, but also a certain amount of translation invariance, which guarantees that the model can identify features when they are slightly misaligned within the image.

- Fully Connected Layers: Placed towards the end of the CNN architecture are fully connected layers that perform high level reasoning and classification. These layers learn complex relationships among extracted features by taking the flattened feature map of the earlier layers. The fully connected layers can then put together all the learned patterns to classify the input data into various categories of emotion, such as happiness, sadness or anger, usually with a final prediction using activation functions such as softmax.

3.3.2. LSTM Model

Long Short-Term Memory (LSTM) model is specially made to process sequential or time-series data and therefore it is very applicable to detecting emotion in speech and video sequences. In contrast to conventional neural networks, LSTM networks are able to maintain long-term memory with memory cells and gating mechanisms, which govern information flow. This enables the model to learn temporal relations and contextual variations in the emotional expression, e.g., change of tone, pitch, or facial movement with time. Consequently, the LSTMs have become popular in the application that requires the study of the dynamics of emotions.

3.4. Training Strategy

An emotion recognition model training strategy aims at maximizing performance and avoiding overfitting and generalizing to new data. The Backpropagation is one of the fundamental elements of this process as it is applied in updating the weights of the network through propagating the error backward to prior layers. In the process, the model will calculate the difference between the predicted and actual outputs in a loss function and will then modify the internal parameters to reduce the error. Backpropagation is used with other optimization algorithms, like Gradient Descent, which repeatedly updates the weights of the model as they move in the direction that minimizes the loss function. Stochastic gradient descent (SGD), Adam, and RMSProp are variants that are typically employed to increase the rate of convergence and stability, particularly in deep learning models. Besides optimization, regularization methods are also important to enhance the robustness of the model and avoid overfitting, where the model is effective on the training data, but ineffective on new data. Training techniques like dropout randomly turn off a set of neurons, compelling the network to learn more generalized characteristics. There are also other methods such as the L1 and L2 regularization which introduces loss function terms that are penalty terms to deter overly complex models with heavy weights. Another successful method is early stopping, in which the training is terminated when the validation performance ceases to increase, thus preventing the needless acquisition of noise in the data. Backpropagation, gradient descent optimization, and regularization methods combine to create an effective training approach to guarantee efficient training, enhanced accuracy, and enhanced generalization in emotion recognition systems.

ISSN: 3071-3315
PP: 01-12.

4. RESULTS AND DISCUSSION

4.1. Performance Metrics

To measure the success of an emotion recognition system, a collection of well-defined performance measures is needed to represent various facets of model behavior. Accuracy is one of the most frequently employed measures, indicating the percentage of the correctly predicted emotions of all the predictions that the model makes. Although accuracy is good to give a brief view of the performance, it is not always accurate when the dataset is skewed i.e. one emotion is represented more often than others. Such metrics as precision and recall are more relevant in such situations. Precision concentrates on the quality of the positive predictions, which means the number of emotions of the model that are correct. This is especially applicable in the event that false positives should be minimized like in sensitive uses like in mental health monitoring. Conversely, recall is used to determine the capability of the model to recognize all the cases of a given emotion that are relevant, i.e. whether the system is sensitive enough to pick up real expressions of emotion and not overlook them. High recall model is one that is sure to miss fewer real emotional cases. The F1-score is a unified measure to strike a balance between precision and recall. It is a combination of both precision and recall into a single value, which is more of a balanced judgment, particularly in case of uneven distributions of classes. The value of F1-score is high, which means that the model has a good balance between the number of correct emotions recognized and the number of errors made. All of these metrics give an overall picture of the model performance, which can guide researchers and developers to comprehend the strengths and weaknesses of the model. With all these measures taken into consideration, it is possible to make sure that the emotion recognition system will not only be accurate but reliable, as well as effective in the real-world setting.

4.2. Experimental Results

Table 1: Experimental Results

Model	Accuracy
CNN	91%
LSTM	88%
CNN + LSTM	93%
DBN	85%

4.2.1. CNN – 91%

Convolutional Neural Networks (CNN) model was able to recognize facial emotions with an accuracy of 91 percent on the FER2013 dataset, indicating a high level of performance. This high accuracy is mainly attributed to the fact that the CNN directly learns the spatial features like facial structures and expression patterns, through images. The FER2013 data has many labeled facial images and they have various emotional expressions, making the model very general. Nevertheless, the lighting, occlusion, and face rotation could also influence the performance.

ISSN: 3071-3315
PP: 01-12.

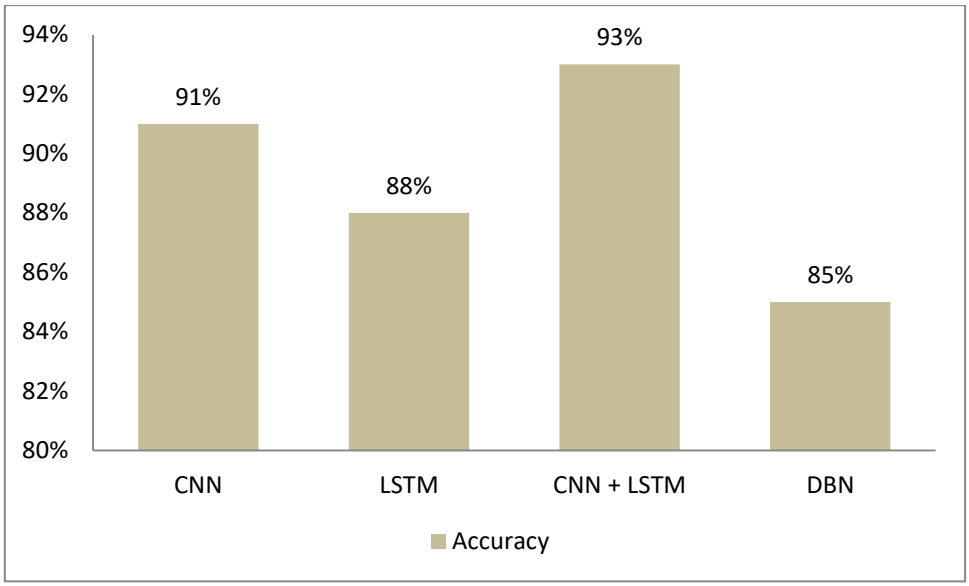

Fig 4 - Experimental Results

4.2.2. LSTM – 88%

Long Short-Term Memory (LSTM) model scored 88 percent on the RAVDESS dataset, which is composed of emotional speech and song recordings. This task is one in which LSTM networks can excel, as they have the ability to learn the temporal relationships in sequential audio information, including tone, pitch, and rhythm. The fact that the accuracy is slightly lower than that of CNN models could be attributed to the complexity of the speech signals and the different vocal expression with diverse speakers.

4.2.3. CNN + LSTM – 93%

The hybrid CNN + LSTM model had the highest accuracy of 93% when it was implemented with a multimodal dataset. In this module, spatial features of visual data are obtained with the help of CNN, and time-related information on the sequences (video frames or audio signals) are processed with LSTM. The system integrates complementary information of the various modalities by using both models to achieve better performance of emotion recognition. This indicates that visual and temporal integration is beneficial in making more accurate and stronger predictions.

4.2.4. DBN – 85%

Deep Belief Networks (DBN) model obtained an accuracy of 85% on the CK+ dataset. DBNs were one of the older deep learning techniques employed in recognizing emotions and they are based on layered probabilistic models to learn feature representations. Although they are quite reasonable in performance, they usually have lower accuracy than more modern architectures such as CNNs, primarily because of scaling and efficiency limitations.

4.3. Analysis

The outcomes of the experiments outline specific advantages of various models to be used in the tasks of recognizing emotions. Facial data is particularly well suited to Convolutional Neural

ISSN: 3071-3315
PP: 01-12.

Networks (CNNs) because they are able to extract spatial hierarchies and visual patterns in images. Face expression typically contains small changes in features like the eye, mouth and the eyebrows, and CNNs are especially capable of learning such local changes by use of convolution and pooling functions. This renders them very appropriate to image based datasets where they could be very accurate even in changing conditions such as lighting and orientation.

Conversely, Long Short-Term Memory (LSTM) networks are much better at modeling time-dependency, since they capture time-related dependencies. Emotions displayed either in speech or video are not fixed, but change gradually through tone, pitch and consecutive movement of the face. LSTMs overcome this through a memory of past inputs, which enables the model to learn the context and flow. This allows them to be very useful with sequential data, but not necessarily as effective as CNNs with solely static images. It is also observed in the analysis that multimodal approaches will produce the most accurate results since they are a combination of the two spatial and temporal models. Multimodal systems offer a deeper insight into what human emotions are by combining the information of visual, audio, and in some cases, physiological signals. This combination decreases uncertainty and increases strength particularly in practical situations where utilizing one source of information might be inadequate. The findings in general indicate that although each of the individual models is beneficial, the integration of multiple modalities results in more precise and valid emotion recognition systems.

5. CONCLUSION

In this paper, a thorough review of the deep learning methods in real-time emotion recognition was provided before 2019, where the focus was on the shift towards more sophisticated data-driven models, replacing the more traditional handcrafted feature-based methods. The previous techniques were very much relying on manually developed features, which were usually restrictive in generalizing to various conditions. The development of deep learning led to the creation of much more accurate and robust emotion recognition systems that include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTMs). CNNs were found to be very useful in deriving spatial feature of face images and RNNs and LSTM were able to improve the potential of the system to capture the temporal variation in speech and video analysis. With these improvements, and the access to larger and more varied data, systems could attain higher levels of performance and generalization.

Although such improvements have been made, there are still a few challenges that curtail the effectiveness of real-time emotion recognition systems. A key problem is that deep learning models are computationally complex, thus challenging to deploy in real-time, particularly on systems with resource constraints, such as mobile phones or embedded systems. Moreover, the expression of emotions differs greatly among people based on the cultural, social and psychological aspects thus making it difficult to have universalized models. Accurate emotion detection is also complicated by environmental noise, as brightness in pictures, or noise in the audio background. The above problems prove that more efficient, flexible, and sturdy solutions are needed.

ISSN: 3071-3315
PP: 01-12.

Future studies should aim at coming up with light and optimized models that can work effectively in real-time without affecting accuracy. Model compression, pruning and quantization techniques may be significant towards this objective. The other significant trend is the direction of cross-cultural emotion recognition, where systems are trained to recognize and respond to the emotional expressions of different populations and enhance inclusivity and reliability. In addition, it is crucial to involve Explainable Artificial Intelligence (Explainable AI) in the systems of emotion detection to enhance transparency and trust. Model decisions will be more interpretable, which allows researchers and users to gain a better insight into the way emotions are being categorized resulting in more ethical and reliable uses. In general, further progress in these directions will help to create more efficient and useful systems of emotion recognition in the future.

REFERENCES

[1] Paul Ekman, "An argument for basic emotions," Cognition and Emotion, 1992.
[2] Timo Ojala et al., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE TPAMI, 2002.
[3] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," CVPR, 2005.
[4] Gajula, S. (2024). Cybersecurity risk prediction using graph neural networks. *Journal of Information Systems Engineering and Management*.
[5] Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," IEEE TASSP, 1980.
[6] Geoffrey Hinton et al., "A fast learning algorithm for deep belief nets," Neural Computation, 2006.
[7] Gajula, S. (2023). A review of anomaly identification in finance frauds using machine learning system. *International Journal of Current Engineering and Technology, 13*(6), 568–575. https://ijcet.evegenis.org/index.php/ijcet/article/view/820
[8] Yann LeCun et al., "Gradient-based learning applied to document recognition," Proceedings of the IEEE, 1998.
[9] Alex Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," NIPS, 2012.
[10] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," Neural Computation, 1997.
[11] Jeffrey Elman, "Finding structure in time," Cognitive Science, 1990.
[12] Björn Schuller et al., "Automatic recognition of emotion-related user states in spontaneous children's speech," IEEE TASLP, 2011.
[13] Gajula, S. (2024). Adaptive zero trust architecture for securing financial microservices. Computer Fraud & Security, 2024(12), 643–655. https://doi.org/10.52710/cfs.845
[14] Maja Pantic and Leon J. M. Rothkrantz, "Toward an affect-sensitive multimodal human–computer interaction," Proceedings of the IEEE, 2003.
[15] Zeng Zhihong et al., "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE TPAMI, 2009.
[16] Sayan Saha et al., "A study on emotion recognition from speech using deep learning," 2018 International Conference, 2018.
[17] Shizhe Chen et al., "Multimodal emotion recognition with deep learning," ACM Multimedia, 2017.
[18] Soujanya Poria et al., "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intelligent Systems, 2017.