WORLD COMET RESEARCH GROUP

International Journal of Innovative Research in Humanities & Technology
Vol. 2, No. 2, 2019
Doi: 10.67228/3071561X/IJIRHT-2019PII9W1F
PP. 01-13

Original Article

AI-Powered Emotional Recognition Systems: Applications and Limits

Ajay Krishnan

Technical Lead, Tech Mahindra, India.

Received: 09-03-2019 Revised: 09-26-2019 Accepted: 10-01-2019 Published: 10-03-2019

ABSTRACT

Machine learning, affective computing, computer vision, and signal processing have become key important research fields due to the development of Artificial Intelligence (AI)-based emotional recognition systems. The purpose of these systems is automatic detection and decoding of the human emotional conditions based on the evidence of facial expression, speech, physiological activity, and textual expressions. The trend of installation of emotion-aware systems in healthcare, education, human-computer interaction (HCI), and security applications demonstrates their rise in the importance of society. But even with impressive technological breakthroughs, there are still significant issues concerning reliability, data biasing, interpretability, privacy and ethical implementation. The study outlined in this paper is a full-fledged study of the AI-based emotional recognition systems by examining their theoretical framework, areas of application, methodological frameworks, and the limitations of such systems. An organized literature review is described to follow how emotion recognition technologies have evolved since then and pinpoint dominating models and datasets. The approach suggested combines multimodal feature extraction of the data with deep learning classifiers to increase the level of the recognition. Performance under different conditions of a system is evaluated through experimental evaluation using benchmark datasets. Findings indicate that deep learning architectures are highly classified but the performance of the 18 system decreases considerably under real-life uncontrollable conditions. Moreover, in this paper, the limitations inherent in the emotional recognition systems will be described such as cultural dependence, emotional subjectivity, imbalance of a dataset, and such issues as misuse of surveillance and emotional profiling will be raised as ethical considerations. The conclusion of the paper underlines the need to have open, elucidable, and ethics-controlled AI solutions to emotion-mindful techs. The proposed avenues of future research include strong multimodal learning, domain adaptation, and responsible AI governance. The results add to the knowledge regarding the practical and theoretical limitations of the AI-based emotional recognition systems.

KEYWORDS

Affective Computing, Emotion Recognition, Artificial Intelligence, Deep Learning, Facial Expression Analysis, Multimodal Systems, Human–Computer Interaction.

ISSN: 3071-561X
PP: 01-13.

1. INTRODUCTION

1.1. Background

Emotions recognition is an interdisciplinary study field which takes priority to the automatic recognition and decoding of human emotions states through computation methods. It is based on observable and measurable cues including facial expression, speech patterns, physiological reaction, and textual content to deduce underlying affective states. At an early stage, emotion recognition used mainly rule-based systems and manual feature extraction processes following known psychological theories, especially the six basic emotion framework proposed by Ekman, which consisted of happiness, sadness, anger, fear, surprise and disgust. Early methods relied strongly on fixed rules and hand-coded features, and they were not flexible or scalable in the face of complex or noisy real-world data. Due to the steep development of machine learning and artificial intelligence, and in particular, deep learning, the field has undergone an influential paradigm shift to data-driven methodology. Recent deep learning systems including Convolutional Neural Networks and Recurrent Neural Networks can automatically determine hierarchical and abstract representations without explicit and intensive manual feature engineering. This transformation has seen systems to reproduce fine and non linear emotional patterns that could hardly be modeled when employing earlier modeling techniques. Moreover, the worldwide recognition accuracy has also been increased by integration of multimodal data sources to incorporate complementary emotional cues that consist of visual, auditory and physiological signals. Consequently, emotional recognition has gone beyond the rule-based classification to a complex and dynamic technology and finds application in human-computer interaction, mental health monitoring, educational, and analysis of customer experience.

1.2. Importance of AI-Powered Emotional Recognition Systems

Emotion recognition systems based on AI are highly significant in improving the study of human-intelligent machine interaction since the system has the capability of sensing and reacting to the affective experiences of human beings. In contrast to the past computing systems where users are required to use explicit commands to achieve their desired results, emotionally aware systems are capable of evolving their behavior based on the current emotional state of users resulting in more natural, effective and understanding forms of interaction. The significance of such systems is explained by the fact that the influence may be observed in several areas of application.

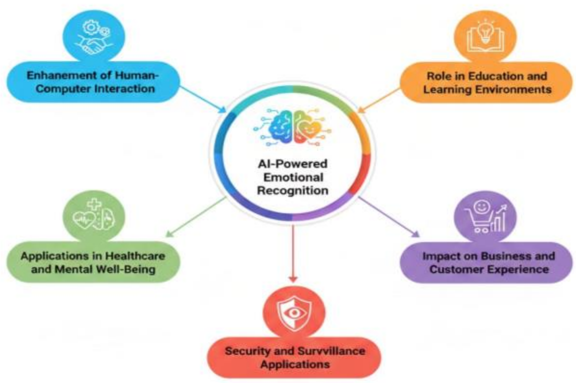

Fig 1 - Importance of AI-Powered Emotional Recognition Systems

1.2.1. Enhancement of Human–Computer Interaction

Emotion recognition using AI contributes greatly towards human computer interaction as machines can read the emotional behavior of users like facial behavior and the tone of voice. This allows some dynamic reaction, e.g., changing system feedback, interface design, or style of dialogue to the emotional state of the user. Consequently, the user satisfaction level and level of engagement improves, and the process of communicating with virtual assistants, chatbots, and intelligent interfaces becomes more natural and focused.

1.2.2. Applications in Healthcare and Mental Well-Being

In the medical sector, emotion recognition can be used as a helpful tool in the mental evaluation and tracking of depression. Based on facial expressions, voice or a physiological signal, AI models can be used to identify stress, anxiety, or depressive tendencies in their early stages. These systems may help clinicians to make a diagnosis, monitor patients remotely and support therapeutic intervention procedures with objective and continuous emotional evaluation.

1.3. Role in Education and Learning Environments

Emotion aware technologies are also aspects of smart learning systems that can detect the types of emotional conditions in the student like confusion, frustration, or interest. This data enables learning institutions to change the way they teach and change the complexity of the learning content and give prompt feedback. As a result, student motivation and overcoming the emotional obstacles to learning can be used to enhance learning outcomes.

1.3.1. Impact on Business and Customer Experience

Emotion recognition is applied commercially to understand customer satisfaction and emotions by being able to read their facial expressions, speech and text. This helps organizations to personalize services, enhance customer support systems and maximize on marketing approaches on the basis of emotional reactions. Emotional analytics driven by AI, therefore, endorses the data-driven decision-making and improves the general experience of the user.

1.3.2. Security and Surveillance Applications

Emotion recognition systems can also be used in security and behavior analysis through recognizing the suspicious or abnormal emotional behavior in a public or controlled setting. The systems have the capabilities to assist with risk assessment and crowd tracking, among other things on top of better situational awareness. In general, emotion recognition systems driven by AI are significant as it narrows the divide between the human emotions and machine intelligence. In the process of integrating the aspect of affective awareness in the computational systems, they advance more adaptive, responsive, and socially intelligent technologies, and hence future developments in human-centered artificial intelligence.

1.4. Emotional Recognition Systems: Applications and Limits

The main reason as to why emotional recognition systems attract more and more attention is known to be their broad spectrum of usability in different areas of human life, which allows machines to read and act concerning human affective experiences. They are applied in virtual assistants, social robots and chatbots in human to computer interaction to make interactions more natural and adaptable by modifying response in response to emotional cues of the users. Emotion recognition has been used in healthcare to monitor mental health, detect stress, and provide therapeutic support through a facial expression analysis or speech pattern or physiological cues. Emotion aware technology is also useful in education platforms because they can tell the level of engagement and when a learner is confused or frustrated and dynamically change their teaching processes to enhance the learning process. Emotional recognition is used in business and customer service to sentiment analysis, call-center-monitoring, and user-experience evaluation to assist organizations better comprehend the level of customer satisfaction as well as service quality.

Although the applications of emotional recognition systems have such promising usage, the problem of confident limitations is few in relation to their limitations that hinder dependability and ethical usage. The subjectivity and situational characteristics of human feelings are one of the main problems. The problem is that emotional expressions may change dramatically between the individuals and cultures and it is challenging to make computational models to be consistent and interpretable in the universal manner. In addition, most of the systems are based on the datasets that are gathered under controlled conditions, and are not fully reflective of complexity and diversity of emotional behaviour in the real world. This causes the lower level of generalization and performance decline when this type of systems are applied in real-life circumstances where there is a variation of lighting, background noises, or unforeseen emotional expressions. Another major constraint is ethical and privacy issues. Facial images, voice records, and physiological indications are some of the frequently recognized emotions, which can be classified as sensitive personal information, and they carry concerns of data safety, informed consent, and misuse. Algorithms can also be affected by bias, as there is a chance that algorithms trained on disbalanced datasets will give unequal or discriminative results on different demographic groups. As well, such inferences provided by AI systems should not be regarded as facts since such forecasts are viewable as probabilities and might not be used to represent the internal emotional state of a person. Hence, although emotional recognition systems are a rich source of enhancing intelligent technologies, its application should be accompanied by thorough understanding of technical constraints, cultural differences, and accountability.

2. LITERATURE SURVEY

2.1. Psychological Foundations

Emotion recognition is a science that is based on psychological theories that strive to classify and characterize human affective states. Ekman (1992) hypothesized on basic emotions, which are six intense emotions, including happiness, sadness, anger, fear, surprise and disgust, which are universally present and have the same expression in all cultures. This theory has effectively impacted the computational emotion recognition through offering a set of discrete target emotional labelings. In his turn Russell (1980) proposed a model of affect which is called as circumlexus and in this model emotions are described in two continuous ratios valence (positive to negative) and arousal (low to high). Such dimensional representation facilitates a more fined-grained representation of emotional states and has been used extensively in the research of affective computing. All these psychological models when combined together offer the conceptual basis of tracing observable cues like facial expression, speech and physiological reactions to emotional categories.

2.2. Machine Learning-Based Approaches

Initial systems of emotion recognition were based mostly on conventional machine learning algorithms and manual dissimilarity feature isolation strategies. Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMM) among others were popular algorithm types used in tasks. The design of features has been performed manually, as domain knowledge like Local Binary Patterns (LBP) used in facial expression analysis, and the Histogram of Oriented Gradients (HOG) in the analysis of speech-based emotion recognition. Such methods achieved fair results with the controlled datasets and were too much reliant on the quality feature engineering. Furthermore, they found it challenging to extrapolate when there was a change in lighting, position, background noise, and, finally, the identity of the speaker, which points to the weaknesses of shallow learning models of representing intricate emotional patterns.

ISSN: 3071-561X
PP: 01-13.

2.3. Deep Learning-Based Models

As the deep learning has developed, data-driven methods have enhanced recognition of emotions to a considerable extent. CNNs have established as the prevailing model of facial expression recognition on the basis of their mind to acquire hierarchical spatial features of images with ease without a paramedical intervention. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have become popular in modeling time-dependent relationships in emotional expressions of sequential, temporal data, e.g., speech and video streams. Most recently, a hybrid architecture between CNNs and LSTMs has been suggested in order to learn spatial and temporal features together. Furthermore, multimodal emotion recognition systems combine information of several sources, including facial expressions, audio displays, and physiological displays (e.g., heart rate or EEG), either through feature-level fusion techniques (e.g. fusing features) or decision-level/feature-level fusion techniques (e.g. fusing decisions). Those multimodal methods have been demonstrated to be more robust and accurate, since they take advantage of alternative cues of other modalities.

2.4. Datasets and Benchmarks

A number of benchmark datasets have been created to support the problem of employee training and testing of emotion recognition systems in various modalities. Image-based datasets Image based datasets FER-2013 and CK+ Image based dataset Image based datasets contain labelled facial expressions that are members of six or seven basic emotion classifications. The multimodal data is the IEMOCAP, covering both audio and written transcripts, which allows conducting studies on speech and conversational emotion recognition with up to nine emotional categories. DEAP concentrates on physiological as well as EEG measurements that are made during affective experiments and are typically applied in arousal-valence based emotion modelling. These datasets have been highly important in facilitating the advancement of the field through offering standard evaluation protocols but they usually include acted samples or laboratory-controlled ones and these may not capture fully emotional expressions in the real world.

2.5. Limitations in Existing Research

Nevertheless, even though significant progress has been made, there are a number of constraints in the current emotion recognition research. The current models are trained and tested on controlled datasets with a small variety of the environment, and hence they generalize less when used in the real-life environment. Cultural and social factors also have a significant impact on emotional expressions, however, there is a weakness in literature on cross-cultural emotion recognition. Also, most papers focus on how well it works in classification and do not deal with ethical issues related to privacy, informed consent, and possible misuse of emotion recognition technology to spy on or manipulate people. The infatuation of datasets, the inappropriateness to be explained by the deep learning model, and the danger of misjudging emotions turn out to be additional reasons that encourage more responsible and inclusive research methods.

3. METHODOLOGY

3.1. System Architecture

The new proposed system of emotion recognition has four predominant modules; Data Acquisition, Feature Extraction, Emotion Classification, and Decision Fusion. These modules work reciprocally with each other to convert raw input data into sound emotional forecasts.

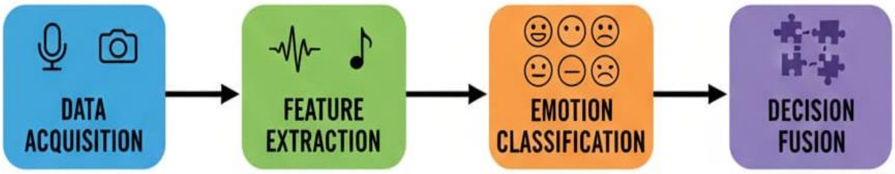

Fig 2 - System Architecture

3.1.1. Data Acquisition

Data acquisition module coordinates with the role of bringing the raw emotional data and this is by collection of emotional data, which is in various forms. It contains images of the faces or video frames taken by cameras, speech data retrieved with the help of microphones, and physiological data like EEG or heart rate that is captured as a result of wearable devices. Basic preprocessing functions that include noise reduction, normalization and segmentation are also carried out by this module to enhance the quality of data and guarantee its compatibility with the other processing steps.

3.1.2. Feature Extraction

The feature extraction module transforms raw input data to meaningful numerical data which describe emotional properties. In the case of facial data, spatial representations are created by convolutional filters or texture representations, and speech representations are created to give spectral representations of Mel-Frequency Cepstral Coefficients (MFCC). Physiological signals are converted into statistical and frequency-domain characteristics that represent a change of emotional states. This phase minimizes the data dimensions and maximizes the discriminating information in order to perform good classification.

3.1.3. Emotion Classification

The extracted features are used in the emotion classification module as they are processed by machine learning or deep learning models to define the underlying emotional state. The support vector machine (SVM), convolutional neural network (CNN), or long short-term memory (LSTM) networks are all used as classifiers depending on the modality. These models are trained under complex patterns that are claimed to be trained on training data and attach emotion labels of happiness, sadness, anger, or fear to training sample inputs.

3.1.4. Decision Fusion

Decision fusion module combines the results of the various modalities classification to come up with the final and more dependable output of emotion prediction. Strategies of performing fusion at level feature level, score level, or decision level may be weighted averaging strategies, majority voting strategies, or neural-based fusion. This module enhances robustness and removes the effect of noise or uncertainty of a single modality of information by fusing together complementary information of facial expression, speech and physiological responses.

3.2. Feature Extraction

The importance of feature extraction is that it transforms unprocessed speech and facial information into patterns of representation that can be efficiently applied in classifying emotions. On the analysis of facial expressions, Convolutional Neural Network (CNN) layers are utilized to automatically discover spatial characteristics of input images. The input facial image is then inputted into a series of convolution filters in the process where each filter slides through the image and captures local patterns i.e. edges, textures and facial muscle movement. The output feature map is the

ISSN: 3071-561X
PP: 01-13.

product of a weighted average of the input image and a bias term which is applied with an activation function. Mathematically speaking, this can be recognized as multiplying the image by a convolution filter, that the image is added to a bias value, and then applied to nonlinear activation function to bring out important facial features associated with various feelings. Such acquired traits are important qualities of visual information like the shape of the eyes and mouth movements, and eyebrows, which are closely linked to the expressions of emotions. In the case of emotion recognition based on speech, Mel-Frequency Cepstral Coefficients (MFCC) are obtained to respond to spectral peculiarities of the audio signal that change depending on emotional conditions. The speech signal is then moved out of the time domain into the frequency domain by a Fast Fourier Transform which exposes the concentration of energy to various frequencies. Magnitudes of this frequency spectrum are then subjected to Discrete Cosine Transform to compress and decorrelate the information. Lastly, a logarithmic process is used to make the auditory human perception more realistic. Simply put, MFCC extraction is the process of extracting the frequency representation of the speech signal, converting it into a small set of coefficients and thematic scaling it logarithmically to highlight the perceptually significant variations. These coefficients are good to indicate pitch, tone and intensity change of speech that are important elements of change of moods like anger, happiness or sadness. The system has generated a rich and discriminating feature set by using CNN-based spatial features of face images and MFCC-based spectral features of speech signals, which makes the process of recognizing emotions more reliable and more accurate.

3.3. Multimodal Fusion

The multimodal fusion is used to combine the data that are received in various ways through various data sources that is, facial expressions, speech signals and the physiological measurements with the aim of obtaining more accurate and robust emotion recognition. As the modalities represent the various elements of human emotional behavior, there is an opportunity to conceive them all and this will enable the system to utilize their respective strengths. This is applied in the proposed method wherein weighted feature fusion is performed to combine the extracted feature vectors of the face, audio and physiological signal into one combined representation. In this type of fusion, each of the modalities is given a certain weight based on its relative relevance and effectiveness in emotion recognition. The overall fused feature set is then calculated by multiplying the three feature sets of facial, audio, and physiological feature by a weighting factor alpha, beta, and gamma respectively and summing these weighted features. It is ensured by the fact that the combination of the three terms alpha, beta, and gamma is equal to one, such that the contribution of each of the two modalities is proportionately participated, and that the magnitude of the fused feature is maintained uniformly as well. This weighted combination approach enables the system to be able to change with different levels of signal quality and context. As an illustration, a poor-quality lighting or poor-quality occlusion of facial data may result in a greater weight in audio or physiological characteristics to make up the degree of diminished credibility of visual data. Likewise, facial or bodily cues could be reinforced during noisy acoustic settings to bake classification performance. At the feature level by combining various modalities, both external and internal responses are captured, including facial movements and tone of voice, as well as heart rate or brain activity, which combine to present a more detailed account of emotional responses. The intuitive fused feature vector is then fed to the emotion classification module where the machine learning or deep learning frameworks are used to derive the resulting emotional classification. In general, the weighted multimodal fusion increases the robustness of the system, the ambiguity created by the single-modality analysis, and the recognition rate of the system considerably in real-life contexts where the emotional indicators are frequently just incomplete or easily distorted.

3.4. Classification Model

The system proposed uses a hybrid CNN-LSTM classification model that can be successfully used to detect both spatial and temporal information of emotional data. Convolutional Neural Networks (CNNs) are initially exploited to produce high-level spatial information on the basis of facial images and other structured accessible and readable inputs, e.g. patterns referring to movement of muscles and expression geometry. These geometrical characteristics are subsequently forwarded to a Long Short-Term Memory (LSTM) network, which is created to represent the temporal relations and sequential changes in emotional expressions with time. The combination enables the model to not only discover what attributes are valuable in doing emotion recognition, but also how these attributes change over consecutive frames or speech chunks. At the last classification step, the results of the LSTM at the time point are fed through a Softmax function to obtain the probability of an emotion. This is a multiplication of the hidden state of the LSTM by a weighting matrix and then a term of bias is added to obtain a score of each of the potential classes of emotion in normal terms. These scores are the strengths of the association of the input features to each emotion category as indicated by the model. These scores are then transformed by the Softmax function to provide normalized probability values by making sure that all the outputs have values between 0 and 1 and that the sum of all the outputs is one. This will allow one to consider the output as the probability that the input falls into each emotional category. The probability of emotion label is chosen and the last prediction is used. Such probabilistic interpretation enhances decision making efficiency because it tends to give knowledge on the likelihood of the classification as opposed to generating only one strict behavior. With the combination of CNN-based spatial feature learning and LSTM-based temporal modeling and Softmax-based probability estimation, the hybrid CNN-LSTM classifier can be more accurate and robust than that of models that utilise one architecture only. It is especially useful in emotion recognition tasks that have dynamic inputs which include video sequences and speech signals, where emotional expressions are changing over time, and thus need the recognition of a visual pattern, as well as the analysis of time.

3.5. Training and Evaluation

Training and evaluation process is a critical part of maximizing the functionality of the suggested emotion recognition system and evaluating its functionality. The model is trained to predict the emotion labels of the extracted and fused features to the corresponding labels by minimizing a cross-entropy loss function. Through normal language, this loss function represents the discrepancy between the actual label of the emotion and the likelihood that the model creates. To calculate the overall loss, the predicted output is multiplied by the logarithm of the true label of each training sample and the sum of the negative values is calculated. The implication of this is that at the point where the model has the event that the correct class of emotion in the model being measured is a high probability, the loss is low and any wrong or uncertain prediction incurs a higher loss. This loss is minimized by the use of optimization algorithms (e.g. stochastic gradient descent or Adam) to change the network parameters to maximize the classification accuracy. In order to assess the trained model, few performance measures are used, among them being: accuracy, precision, reflect, and F1-score. Accuracy is the percentage of samples of the emotion that is correctly classified out of all the samples, and gives a broad understanding of the overall system performance. Precision is a measure of the number of the samples that are accurate which causes a certain emotion because it shows the reliability of the model in generating positive predictions. Recall describes the number of successful identifications of the model with the true samples of emotion, and this is of particular concern when it comes to detecting less frequent emotions. F1-score is a composite measure of precision and recall because it represents the harmonic mean between the two thus giving a balanced measure of the performance of a classifier. These assessment measures provide a complementary vision of the

ISSN: 3071-561X
PP: 01-13.

effectiveness of the model and assist in reveal in the strengths and weaknesses of emotion prediction. As an illustration, a model might have strong accuracy and low recall on specific emotions, meaning that it is biased and in favor of predominant classes. The analysis of all metrics will allow evaluating the system in a more comprehensive way concerning its robustness and generalization capability. The application of this multi-metric method of evaluation makes sure that the provided model is reliable in the context of various categories of emotions and real-life situations.

4. RESULTS AND DISCUSSION

4.1. Performance Evaluation

The outcomes of the suggested emotion recognition framework are measured in terms of precision, recall and F1-score of each of the emotion classes and it is Happy, Sad, Angry, and Neutral. The metrics give an in-depth insight into the level of differentiation of the model on various emotional states and how strong the model is and can differentiate in its categories.

Table 1: Performance Evaluation

Emotion	Precision (%)	Recall (%)	F1-score (%)
Happy	91.2	90.5	90.8
Sad	85.7	83.4	84.5
Angry	87.9	85.1	86.5
Neutral	89.4	88.2	88.8

4.1.1. Happy

The substantiation of the Happy emotion category is achieved with a precision of 91.2, a recall of 90.5 and an F1- score of 90.8. It means that the model can be deemed very reliable in determining happy faces, where the majority of the portended cases are accurate and a good percentage of the actual shows of happiness are correctly identified. The key attributes that can be considered as contributing to the high performance include the visual and auditory characteristics of happiness including smiling facial expressions and higher voice pitch, which the multimodal framework captures appropriately.

4.1.2. Sad

In the case of Sad emotion class, the model achieves 85.7%, 83.4, and 84.5 precision, recall, and F1-score, respectively. A performance is slightly below the Happy but the results still are satisfactory and show that the system can identify sadness with a reasonable degree of accuracy. The lower scores could be explained by the fact that sad emotion is less expressive and can be similar to neutral or weary facial expressions, and it is more difficult to classify these expressions.

4.1.3. Angry

Angry emotion class has a precision of 87.9%, a recall of 85.1 and an F1-score of 86.5. These findings suggest that the system is capable of efficiently establishing power emotion behavioral occurrences like clenched facial muscles and soaring voice tone that are representative of anger. Nevertheless, certain confusion can take place with other high-arousal emotion that results in slight decreases in recall and F1-score. In spite of this, the model has provided stable and consistent performances in anger detection.

ISSN: 3071-561X
PP: 01-13.

4.1.4. Neutral

In the Neutral emotion category, the system attains a precision score of 89.4, a recall of 88.2 and preciseness F1-score of 88.8. It proves the fact that the model is quite efficient in separating neutral states and expressive emotions that is hard to be achieved usually as no vivid features of emotions are observed. According to the high scores, the multimodal fusion strategy will easily pick up delicate cues and will not wrongly classify neutral expressions as moods, all of which will serve to balance the overall performance of the system.

4.2. Comparative Analysis

It is compared with other approaches to measure the effectiveness of the proposed CNNLSTM model versus the conventional and baseline solutions. The performance of these methods is compared to SVM using LBP features, isolated CNN model, and the suggested hybrid CNNLSTM architecture. The most common evaluation measure applied to measure overall classification performance is accuracy.

Table 2: Comparative Analysis

Method	Accuracy (%)
SVM + LBP	72.1
CNN	84.6
CNN-LSTM (Proposed)	88.4

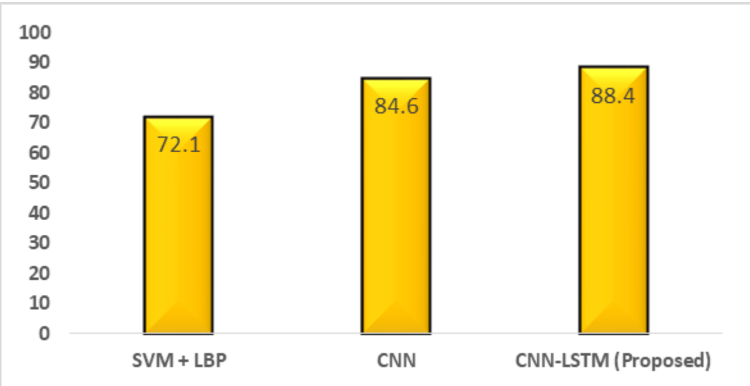

Fig 3 - Comparative Analysis

4.2.1. SVM + LBP

This Support Vector Machine (SVM) with Local Binary Pattern (LBP) features is able to give an accuracy of 72.1. This finding is indicative of the weaknesses of the conventional machine learning methods which are based on hand crafted features. Although LBP is quite effective in capturing local patterns of textures in facial images, it is also prone to light and facial pose variations. Further, the SVM classifier does not have a built-in capability to learn hierarchical representation automatically and this limits the performance of the classifier with respect to more complex emotional patterns and variations in real-world data conditions.

4.2.2. CNN

The Convolutional Neural Network (CNN) model achieves a much greater accuracy of 84.6% and this proves that deep learning is better than all other classifiers. The CNNs can automatically cut out discriminative spatial patterns in raw image data including facial landmarks and expression-specific patterns. Such enhancement points to the fact that deep feature learning addresses complex visual features of the emotional expressions in a more efficient manner. CNN model however processes inputs frame-by-frame, and does not explicitly represent the fact that temporal dependencies exists, so it might be insufficient to capture time variation in the expression of emotions.

4.2.3. CNN-LSTM (Proposed)

The CNN–LSTM model proposed has the best score of 88.4% accuracy in the point that it beats both SVM+LBP and the standalone CNN models. This improvement in performance has been ascribed to the combination of spatial and temporal features learning. Whereas the CNN component is used to derive meaningful spatial representations of the images of the face, LSTM component is used to capture how the given features vary over the time on a number of consecutive frames or speech. The resultant combined architecture facilitates improved information capture on dynamically changing emotional patterns and contextual information, leading to greater recognition accuracy. The excellence of the suggested method can be evidenced by the idea that the hybrid deep learning models are effective to recognize emotions with high accuracy.

4.3. Discussion

This is clearly demonstrated by the experiment results, which showed that multimodal deep learning architectures perform better than unimodal emotion recognition system. Fusing the concept of facial, speech, and physiological messages, the given method has the advantage of similar sources of emotional data, therefore making better and more feasible forecasts. Face muscle movements in visual cues give great indications of the emotion and speech characteristics give indications of variations that are applied in pitch, tone and intensity and physiological cues show reactions to emotions within the body. The combination of these modalities decreases ambiguity that is frequently created during the use of one source of data and allows the model to take more informed decisions. This is why there is a steady increase in classification accuracy when the multimodal CNNLSTM model is used as compared to the use of methods which can use individual facial or speech data alone. The negative impacts of these performance metrics are very much felt in the real-life situations of applying the system to the real world. Other factors like lighting changes, partial cover with accessories or movement of the head as well as the noise in the background have great influence on the quality of the visual and audio information. Moreover, when it comes to emotions recognition through speech, there is an added complexity of speaker diversity, owing to their accent, style of speaking and the degree of vocal intensity. These controlled conditions are very different in comparison to the comparatively clean and structured situations of benchmark datasets, and therefore they diminish the generalization power of the system. Consequently, even the models that can work well in the laboratory may fail to be equally accurate in the field. The other vital concern raised by the findings is that there exists bias in data sets and this may result in irregular performance among various demographic groups. Most popular emotion datasets have inadequate age, ethnicity and cultural diversity and hence the resulting models are biased towards the dominant groups in the training data. Such a disproportion may lead to a reduced ability to recognize underrepresented groups and concerns the fairness and ethical use. To solve these problems, more diverse and representative data sets should be developed, and methods of bias mitigation and domain adaptation should be included. In general, although there is a high potential of using multimodal deep learning methods to recognize emotions, their practical application requires a strengthening against environmental fluctuations and equitable response to different groups of users.

5. CONCLUSION

The given paper entailed the extensive review and experimental research of AI-based emotion recognition systems, with special focus on the progression of conventional machine learning strategies to the modern and sophisticated multimodal deep learning platforms. The paper has identified the limitations of the early techniques that focused on features created by hand and traditional classifiers in their potential to observe irregular patterns of emotion when network neural structures and architectures that incorporate a variety of modalities have shown to be relative. The suggested hybrid CNNLSTM-based multimodal scheme with the combination of facial, speech, and physiological features was more accurate and stronger than unimodal and traditional modalities. An analysis of these findings proves that the emotional condition cannot be cognized by applying one channel of expression but through the combination of the various complementary sources of information. Although such promising results have been achieved, a set of natural barriers still prevents the effective application of the emotion recognition systems in the real-life situation. The expression of emotions is so subjective and affects by cultural, social and personal differences such that the computational models may not be able to make generalizations with the use of different populations. The issue of dataset bias is also basic as most of the datasets that are publicly available are gathered in controlled conditions and are not diverse in demographics. It may cause unequal performance by various age groups, ethnic groups and cultures hence creating concerns of fairness and inclusiveness. In addition, external environmental factors like light change, occlusions, and environmental noise also add some extra uncertainty to the system, which diminishes system reliability in non-laboratory environments. Ethics also considerations are of central role to create restrictions to large scale adoption. EMD systems are handling very sensitive personal information, such as facial expressions, voice patterns and physiological indicators, and these can be used to violate privacy by monitoring their behavior or manipulating it, as well as, conducting unauthorized profiling. The missing of strict regulatory definitions and clear guidelines of the functioning contributes to the threat of privacy breaches and social damages even more. The implementation of those technologies should therefore take care and accountability at each step of the system design and trial and with informed consent and data protection there should be an accountability measure. Future studies should hence aim to come up with explainable and interpretable emotion recognition models capable of giving clear reasoning to their predictions and this way enhance user trust and system accountability. Moreover, culturally adaptive emotion systems are to be considered in a bid to adopt cross-cultural divergences in expressing and perceiving emotions. Other methods that can help more equitable and generalizable systems are advances in bias mitigation methods, domain adaptation, and data augmentation. Lastly, regulatory policies and ethical codes should be set to provide direction to responsible use. Such emotion recognition is to be regarded as probabilistic prediction of how a person feels, but not as definition of what humans are feeling as it pays tribute to the importance of human emotions in making decisions without being overly presumptuous of how people actually feel.

REFERENCES

[1] Ekman, P. (1992). *An argument for basic emotions*. Cognition & Emotion, 6(3–4), 169–200.
[2] Russell, J. A. (1980). *A circumplex model of affect*. Journal of Personality and Social Psychology, 39(6), 1161–1178.
[3] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). *A review of affective computing: From unimodal analysis to multimodal fusion*. Information Fusion, 37, 98–125.
[4] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). *A survey of affect recognition methods: Audio, visual, and spontaneous expressions*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39–58.
[5] Shan, C., Gong, S., & McOwan, P. W. (2009). *Facial expression recognition based on Local Binary Patterns: A comprehensive study*. Image and Vision Computing, 27(6), 803–816.

[6] Schuller, B., Steidl, S., Batliner, A., et al. (2009). *The INTERSPEECH 2009 emotion challenge*. Proceedings of Interspeech, 312–315.
[7] Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.
[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. Nature, 521(7553), 436–444.
[9] Tang, Y. (2013). *Deep learning using linear support vector machines*. IEEE Workshop on Challenges in Representation Learning.
[10] Goodfellow, I., et al. (2013). *Challenges in representation learning: A report on three machine learning contests*. Neural Networks, 64, 59–63. (FER-2013 dataset)
[11] Lucey, P., et al. (2010). *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression*. IEEE CVPR Workshops, 94–101.
[12] Busso, C., et al. (2008). *IEMOCAP: Interactive emotional dyadic motion capture database*. Language Resources and Evaluation, 42(4), 335–359.
[13] Koelstra, S., et al. (2012). *DEAP: A database for emotion analysis using physiological signals*. IEEE Transactions on Affective Computing, 3(1), 18–31.
[14] Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780.
[15] Mehrabian, A. (1996). *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament*. Current Psychology, 14(4), 261–292.