Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 21-40

Affective Gesture Recognition in Virtual Reality Using LSTM-CNN Fusion for Emotion-Adaptive Interaction

Soonya Gupta*[1], Deepa Kumar[1], Shiva Sharma[1]
Email: soo.gupta@igdtuw.ac.in; deepakumar@igdtuw.ac.in; shiva112@igdtuw.ac.in
[1]*Department of Information Technology, Netaji Subhas University of Technology (Formerly Netaji Subhas Institute of Technology), New Delhi, India*
*Corresponding Author

Abstract
Emotion recognition in Virtual Reality (VR) has become increasingly relevant for enhancing immersive user experiences and enabling emotionally responsive interactions. Traditional approaches that rely on facial expressions or vocal cues often DECIA limitations in VR environments due to occlusion by head-mounted displays and restricted audio inputs. This study aims to develop an emotion recognition model based on body gestures using a hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). The CNN component extracts spatial features from skeletal data, while the LSTM processes the temporal dynamics of the gestures. The proposed model was trained and evaluated using a benchmark VR gesture-emotion dataset annotated with five distinct emotional states: happy, sad, angry, neutral, and surprised. Experimental results show that the CNN-LSTM model achieved an overall accuracy of 89.4%, with precision and recall scores of 88.7% and 87.9%, respectively. These findings demonstrate the model's ability to generalize across various gesture patterns with high reliability. The integration of spatial and temporal features proves effective in capturing subtle emotional expressions conveyed through movement. The contribution of this research lies in offering a robust and non-intrusive method for emotion detection tailored to immersive VR settings. The model opens potential applications in virtual therapy, training simulations, and affective gaming, where real-time emotional feedback can significantly enhance system adaptiveness and user engagement. Future work will explore real-time implementation, multimodal sensor fusion, and advanced architectures, such as attention mechanisms for further performance improvements.

Keywords: *Emotion Recognition, VR, Body Gestures, CNN-LSTM, Affective Computing.*

I. INTRODUCTION

In the rapidly advancing era of immersive technology, VR has emerged as one of the primary mediums in Human-Computer Interaction (HCI). With its ability to create immersive and interactive environments, VR offers significant opportunities across various domains, such as education, training, entertainment, and therapy. However, a key challenge in creating truly effective VR experiences lies in how systems can respond to and adapt to users' emotional states in real time. Emotionally adaptive interaction is essential for fostering more natural, personalized, and human-centered VR experiences (Arman et al., 2022). One promising approach to achieving this goal is through the utilization of affective computing based on non-verbal modalities, particularly body gestures. As a universal form of non-verbal communication, gestures carry a wealth of emotional information that can be used to understand an individual's affective state.

Received on January 2025; Revised on February 2025; Accepted on March 2025; Published on April 2025
DOI: *10.51903/jtie.v4i1.278*

With the advancement of artificial intelligence technologies, particularly in the field of deep learning, the CNN and LSTM approaches have demonstrated promising results in simultaneously processing spatial and temporal data. CNNs are known for their effectiveness in extracting visual features from spatial data such as images or video frames, while LSTMs excel in recognizing sequence patterns and temporal dynamics. According to (Huang et al., 2023), the combination of CNN and LSTM in emotional gesture recognition in video has achieved high accuracy rates of up to 84%. Meanwhile, (Swoboda et al., 2022) employed a Support Vector Machine (SVM)-based approach with pose estimation to recognize static gestures; however, this approach does not fully accommodate temporal information. On the other hand, (Kaklauskas et al., 2022) highlighted that research in affective computing tends to focus more on facial expressions and voice as the primary indicators of emotion. A similar observation was made by (Atmaja et al., 2022), who stated that face- and voice-based approaches still dominate in affective adaptive systems.

Although numerous studies have been conducted in the field of affective computing within VR environments, most developed approaches remain focused on facial expressions and vocal signals as the main sources for detecting user emotions. For instance, (Izountar et al., 2022) developed an emotion detection system based on facial expressions in VR simulations for medical training, while (Dirin & Laine, 2023) employed voice analysis as an emotional indicator in VR-based educational games. Conversely, research by (Lian et al., 2023) demonstrated that the integration of voice and facial expressions can enhance emotion recognition accuracy, yet it still overlooks the potential of body gestures. According to (Siddiqui et al., 2022), body gestures contain rich emotional information and can serve as a complementary modality in adaptive systems; however, their study has yet to integrate deep learning approaches in depth. Additionally, Rani and (Rani & Devarakonda, 2022) explored the use of CNN-LSTM for recognizing emotions from gesture videos, but this approach was not implemented within a VR context that requires real-time adaptive systems. Therefore, this study aims to address this gap by developing an effective gesture classification system based on CNN-LSTM fusion within a VR environment to enable more adaptive and context-aware emotional interactions.

Through this approach, the study seeks to build an emotional gesture recognition system that is not only accurate but also applicable in real-time interactive applications within VR environments. The central hypothesis of this research is that the combination of CNN and LSTM can effectively capture the spatio-temporal patterns of users' body gestures and classify them into relevant emotional categories. The utilization of this deep learning architecture is expected to optimize both the accuracy and efficiency of the system in dynamically recognizing affective expressions, particularly in gesture-based interactions. Furthermore, the development of this system also takes into account the need for adaptive responses to users' emotional changes in

complex VR simulations. This research is not only focused on emotion recognition alone but also on integrating the system into a VR framework that enables behavioral adaptation of the virtual environment based on emotion classification outcomes. The results of this study are expected to make a significant contribution to the development of more natural and immersive affective interaction technologies while also expanding the scope of multimodal emotion recognition research within immersive virtual ecosystems.

II. LITERATURE REVIEW

A. Theoretical Foundations

1. Affective Computing in the Context of Virtual Interaction

Affective computing is a multidisciplinary field aimed at enabling computational systems to understand, recognize, and even express human emotions through various sensory modalities. In the context of virtual interaction, affective computing plays a crucial role in bridging the emotional gap between users and digital environments, thereby fostering more human-centered and contextually relevant experiences. According to (Mourtzis et al., 2023), affective systems can enhance the quality of human-machine interaction by adjusting responses based on emotional signals detected from users. This approach utilizes a range of non-verbal inputs, such as facial expressions, vocal intonation, and body gestures, to identify emotional states in real time. The technology has been widely implemented in various applications, including simulated training, virtual education, and technology-based healthcare, where emotional awareness plays a vital role in the effectiveness of interactions. The ability of systems to detect and interpret emotional signals enables deeper interactive experiences, particularly within complex digital environments.

As sensor technologies and data processing capabilities continue to advance, affective computing within VR environments is increasingly geared toward capturing and analyzing affective behavior through the integration of multimodal information sources. Researchers such as (Grewal et al., 2022) emphasize that in digital interaction, users' emotional responses are often not fully conveyed through verbal communication alone, necessitating approaches that incorporate physiological signals and other non-verbal behavioral cues. The integration of VR with affective systems allows virtual environments to respond in more natural and context-aware ways—for instance, through changes in visual ambiance or the dynamics of virtual character interactions. Such interaction enriches the user experience by providing responses aligned with the observed emotional states. These systems also enable virtual environments to adapt to users' emotional reactions without explicit intervention, creating a more personalized and contextually relevant form of interaction based on affective data collected throughout the process.

In its implementation, affective computing requires computationally processable representations of emotional data, whether derived from facial expressions, vocal patterns, or body movements. As explained by (Udahemuka et al., 2024) in the DEAP study, multimodal approaches to emotion recognition yield higher performance as they are capable of capturing emotional nuances from multiple perspectives. Immersive VR environments provide a rich context for emotion recognition systems, allowing for more accurate mapping between affective cues and virtual scenarios. The use of physiological data such as Electroencephalography (EEG), heart rate, and facial muscle activity within VR is increasingly being explored to enhance the accuracy of affect detection. At the same time, behavioral signals such as body gestures are gaining attention as practical and non-invasive indicators of emotional expression. The integration of physiological and gestural data opens new possibilities for a more holistic understanding of users' emotional dynamics.

In the realm of virtual interaction, researchers such as (Yuvaraj et al., 2025) have demonstrated that the integration of affective computing can improve system responsiveness to users' affective changes, particularly in simulation-based training and therapeutic contexts. They argue that VR systems capable of detecting and responding to users' emotions have the potential to enhance the quality of simulations in terms of both realism and emotional engagement. Such systems can be adapted based on affective data to generate dynamic and psychologically relevant scenarios. In practice, this facilitates the creation of more immersive interactions by accommodating users' continuously evolving emotional needs. The application of affective computing also supports the development of virtual avatars or agents that appear more empathetic and emotionally responsive in technology-mediated social interactions. This affective component becomes a critical element in shaping digital experiences that are more immersive and closely resemble real-world interpersonal interactions.

2. Gesture Recognition in the Context of Human-Computer Interaction

Gesture recognition has emerged as a pivotal field within HCI, as it enables users to interact with systems naturally through body movements without relying on traditional physical input devices. According to (Strazdas et al., 2022), body gestures can serve as an intuitive means of communication that closely resembles human-to-human interaction, thereby enhancing user engagement in interactive systems. Gesture recognition systems typically rely on vision-based techniques such as image processing and skeletal tracking to identify movement within a three-dimensional space. This technology has been increasingly applied in a variety of domains, including digital gaming, interactive presentation systems, and motion-sensor-based medical rehabilitation. One of the defining characteristics of gesture-based interaction is its ability to

directly capture physical expression, which is often absent from text- or voice-based interfaces. The growing use of gesture-driven approaches has, in turn, stimulated further exploration within the realm of HCI.

Over time, gesture recognition approaches have shifted toward leveraging machine learning models to improve the accuracy and generalizability of interactive systems. According to (Rahman et al., 2024), machine learning techniques such as SVM, k-Nearest Neighbor (k-NN), and Hidden Markov Model (HMM) have made significant contributions to detecting movement patterns from video or sensor data. These models are trained to recognize both static and dynamic gestures based on visual features or sequences of user skeletal position data. Beyond technical considerations, contextual factors in interaction, such as movement speed, direction, and intensity, have become central to the development of gesture-based HCI systems. These data are often captured using sensors like Kinect, Leap Motion, or RGB-D cameras, which are capable of acquiring spatial information in real time. Effective gesture recognition systems are designed to process data rapidly while maintaining precision in identifying relevant movement patterns.

Gesture recognition in HCI also contributes to enriching affective communication between humans and machines by interpreting bodily expressions that reflect emotional states. As explained by (Khan et al., 2024), body gestures are frequently used to convey emotional nuances such as tension, joy, or discomfort, which can be detected through changes in movement patterns. Such studies combine motion recognition with affective expression analysis to broaden the scope of non-verbal communication in interactive systems. Utilizing gestures as indicators of emotion plays a crucial role in supporting interactive systems that require an understanding of users' psychological states. In immersive environments such as VR, facial expressions are often obscured by head-mounted devices, making body gestures one of the primary modalities for recognizing users' emotions. Emphasizing non-verbal communication through gestures further strengthens the relevance of motion recognition in the design of modern HCI systems.

The integration of gesture recognition in VR and HCI has progressed significantly with the development of deep learning models capable of processing spatio-temporal data simultaneously. According to (Kaseris et al., 2024), the combination of body position data with deep neural networks such as Deep Belief Networks (DBN) and Recurrent Neural Networks (RNN) enables systems to recognize complex movement patterns that unfold over specific time sequences. These models account for the continuity of motion and temporal context, making them more adaptive to shifts in users' interaction patterns. VR environments, which offer users full freedom of movement, generate highly varied gesture data that necessitate sequence-based classification approaches for accurate interpretation. In this context, gesture data may convey control

commands as well as information related to the intensity of emotional expression. The advancement of such methods expands the potential for processing bodily cues in interactive scenarios that rely on understanding users' emotional dynamics and behavioral patterns.

B. Previous Studies

1. Gesture-Based Emotion Recognition Using CNN and LSTM

Research on gesture-based emotion recognition has demonstrated significant potential through the combined application of CNN and LSTM architectures. According to (Leong et al., 2023), the use of these two models enables systems to simultaneously capture spatial and temporal information from video data, particularly in recognizing human emotions through bodily expressions. In their experiments, the system successfully classified six basic emotion categories: anger, happiness, sadness, fear, disgust, and neutral with relatively high accuracy, achieving 78.52% on the CREMA-D dataset and 63.35% on the RAVDESS dataset. CNN was employed to extract visual features from individual video frames, while LSTM was utilized to model the temporal sequence of recorded body movements. This combination allows the system to recognize emotional dynamics occurring over time, rather than relying solely on static images. The implementation of this approach also highlights the necessity of integrating multimodal data in human-machine interaction to facilitate a more comprehensive understanding of affective states.

(Zheng & Blasch, 2023) developed a hybrid model that combines CNN with Convolutional LSTM (ConvLSTM) to improve emotion recognition performance in videos, particularly in facial expressions. They emphasized that the ConvLSTM approach is capable of preserving the spatial structure of image data without transforming it into one-dimensional vectors, as is typical in standard LSTM models. In tests using the SAVEE, CK+, and AFEW datasets, this model achieved competitive results compared to previous methods while maintaining a low parameter count. This demonstrates that spatial information in facial expressions can be better preserved and analyzed using convolutional network structures adapted for sequential data. The researchers also noted that ConvLSTM is more effective in handling noise and unstructured motion variations in real-world videos. The integration of this model expands the possibilities for exploring complex visual data in the context of real-time emotion recognition.

Meanwhile, (Kopalidis et al., 2024) proposed a CNN-LSTM-based approach specifically designed to address the issue of overfitting, which frequently affects conventional CNN models in facial expression recognition. In this model, CNN is used to extract spatial features from each image frame, while LSTM captures the temporal correlations between frames in a sequence. The addition of LSTM after CNN contributes to improved classification accuracy, particularly when

facial expression patterns shift gradually over time. In experiments using the CK+48 dataset, the model achieved an accuracy of up to 84%, demonstrating the effectiveness of combining the two types of features. Liu and his team also noted that the temporal stability of facial expressions is a critical element for consistently detecting emotions. This network structure forms the foundation for the development of interactive systems capable of interpreting users' emotional nuances based on sequential facial and bodily movements.

Other studies have also supported the effectiveness of combining CNN and LSTM in the context of gesture-based emotion recognition, whether from facial or bodily cues. According to (Chouhayebi et al., 2024), the use of spatio-temporal features is essential for understanding the emotional context of movements displayed by users in various interaction scenarios. The model they developed utilizes CNN to capture visual feature representations from body poses, while LSTM processes the sequence of movements occurring throughout the interaction. The study showed that body gestures conveying emotions such as joy or anger can be automatically recognized by learning consistent movement patterns in video data. This approach was also validated in environments with varying lighting and backgrounds, demonstrating the system's adaptability to real-world conditions. The findings from these various studies reinforce the evidence that combining spatial and temporal architectures provides a high representational capacity for comprehensively understanding emotional expressions.

2. Development of Emotion-Based Adaptive Systems in Virtual Reality Environments

Research on the development of emotion-based adaptive systems within VR environments has demonstrated significant potential in enhancing the quality of HCI. According to (Shomoye & Zhao, 2024), automatic emotion recognition in VR classrooms can be achieved through the implementation of CNN, particularly a customized ResNet50 model designed to recognize facial expressions such as being neutral, bored, happy, and confused. This model is engineered to function effectively even when parts of the user's face are obscured by a VR headset, with a primary focus on the lower facial region. This approach lays an essential foundation for VR systems capable of responding in real time to users' emotional dynamics. Moreover, the model enables continuous emotional analysis, which can be applied in various contexts, such as online learning, simulation-based training, and virtual social interactions. The implications of this approach suggest that emotional elements can be effectively integrated into VR system architectures to enhance their sensitivity to users' affective needs.

Other researchers, such as (Vrskova et al., 2023), developed a hybrid architecture combining 3D-CNN and ConvLSTM to perform in-depth analysis of facial expressions from video data. This method leverages CNN's ability to extract spatial features and LSTM's capacity to model

temporal sequences, resulting in a system capable of understanding users' emotional dynamics over time. Unlike traditional LSTM models, this approach preserves spatial information without converting data into one-dimensional vectors, thereby maintaining greater accuracy throughout the training and inference processes. When tested on various datasets, including SAVEE, CK+, and AFEW, the model demonstrated high accuracy and parameter efficiency, making it highly applicable for VR platforms with computational limitations. Beyond accuracy, the model's real-time capabilities further support its suitability for emotion-aware interactive environments. These findings illustrate how CNN and LSTM integration can enhance a system’s comprehension of nonverbal emotional expressions in virtual scenarios.

In addition to facial expression recognition, a study by (Yaseen et al., 2024) examined dynamic hand gesture recognition using a combination of 3D-CNN and LSTM to capture information from sequences of body movements in video. This model was designed to recognize gestures accurately within realistic timeframes, enabling the processing of body signals as indicators of emotional states. The system utilizes spatial features extracted from video frames by CNN and temporal dynamics analyzed by LSTM to form sequential representations of gestures. This approach has proven effective in classifying various types of hand gestures associated with emotions such as anger, happiness, and sadness. The implementation of such a system is particularly relevant in VR contexts, where responsiveness to users' body gestures is a crucial aspect of affective communication. This approach provides a robust foundation for developing gesture-based interactive systems that can interpret the emotional context of user movements.

Furthermore, (Alabdullah et al., 2023) developed a combined CNN and RNN model for dynamic hand gesture recognition using depth and skeletal data. This method aims to extract spatial information accurately from three-dimensional data while capturing the temporal sequence of body movements. By leveraging input from sensors such as Kinect, the model is able to identify detailed changes in posture and movement direction. The use of skeletal data allows the system to interpret gestures based on the relative positions of body parts, which is crucial for distinguishing between different gestures. In their experiments, the system demonstrated robustness in detecting complex gestures that reflect specific emotional states. This approach expands the understanding of how gesture recognition can be applied within virtual environments to build systems that are more responsive to users' non-verbal communication. A comparative summary of the characteristics of various studies on emotion-based adaptive systems in virtual environments is presented in Table 1.

Table 1. Comparison of Studies on the Development of Emotion-Based Adaptive Systems in Virtual Reality Environments

Researchers	Type of Emotional Data	Target VR Implementation	Key Advantages
(Shomoye & Zhao, 2024)	Facial expressions (lower face area)	VR-based virtual classrooms	Detection of emotional expressions despite partial facial occlusion by headset
(Vrskova et al., 2023)	Facial expressions (video)	VR platforms with limited computation	Spatial and temporal preservation; high accuracy
(Yaseen et al., 2024)	Hand gestures (video)	Gesture-based VR interaction	Real-time classification of emotionally expressive gestures
(Alabdullah et al., 2023)	Body gestures (3D skeletal & depth)	Motion recognition systems in VR	High accuracy in recognizing complex gestures based on body part positioning

III. RESEARCH METHOD

This study employs a quantitative experimental approach by utilizing deep learning methods and user experiments within a VR environment. The primary focus of this approach is to evaluate the effectiveness of the model in recognizing emotions based on body gestures systematically and measurably. The system is designed to operate in realistic interaction scenarios by taking into account the temporal aspects of users' gestural expressions. Evaluation is carried out through a series of controlled trials involving real participants in a VR setting, aiming to collect data that reflects actual usage conditions. The results from these tests are used to assess the model's performance and analyze its accuracy in classifying emotions based on bodily movements. This strategy supports the development of an affective system that is adaptive and responsive to users' emotional dynamics in the context of VR-based HCI.

The dataset used in this study is a custom dataset recorded directly from participants wearing VR headsets under controlled conditions. Data collection was conducted by instructing participants to perform various body gestures associated with specific emotions—namely happiness, sadness, anger, and neutrality while remaining within the VR simulation environment. Data capture was carried out simultaneously with body pose tracking to generate video frames that sequentially represent each gesture. Emotion labeling was performed through a combination of researcher observation and participants' subjective self-reports to ensure the validity of emotional interpretations. The data were then organized into time-sequenced frames to be used as input for the deep learning architecture. Information regarding the distribution of data by emotion category recorded in the dataset is presented in Table 2 as part of the dataset documentation.

Table 2. Emotion Class Distribution in the Dataset

Emotion Label	Number of Sessions	Number of Frames
Happy	40	8.000
Sad	38	7.600
Angry	42	8.400
Neutral	40	8.000

The emotion recognition model is built using a combined CNN and LSTM architecture, specifically designed to process spatio-temporal data derived from body gestures. The CNN component functions to extract spatial features from each video frame, such as body contours and limb orientations, to form informative visual representations. These spatial features are then passed into the LSTM network, which is capable of recognizing the temporal sequence of gestures recorded across the frame sequence. The CNN employed in this study is a pre-trained ResNet-18 model that has been adapted to process body image inputs with a resolution aligned to standard training specifications. The integration of CNN and LSTM enables the model to learn the relationship between gesture forms and temporal changes in bodily expressions that convey emotional states. Details of the model architecture and key training parameters used during the training and evaluation process are summarized in Table 3.

Table 3. CNN-LSTM Model Architecture and Training Parameters

Component	Details
CNN Backbone	ResNet-18, pre-trained on ImageNet.
Input Size	224 × 224 × 3
LSTM Layer	2 layers, 256 hidden units
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Epochs	50

The analytical process in this study began with a data preprocessing stage to ensure optimal input quality for model training. This stage included the segmentation of frames based on their temporal sequence within gesture sequences, normalization of image size and intensity to maintain consistency, and data augmentation to increase the diversity of representations without

explicitly adding new data. The augmentation techniques employed included random rotation, horizontal flipping, and brightness adjustment to enhance the model's generalization ability toward variations in real-world data. Following these steps, each image frame was processed by a CNN model to extract significant spatial features, which were subsequently passed as input to the LSTM model to capture the temporal dynamics of the gesture sequences. The trained model then produced emotion labels as predictions, indicating the dominant emotional category expressed in each body movement sequence.

Model performance evaluation was conducted using relevant metrics to assess classification accuracy and prediction error, namely accuracy, F1-score, and the confusion matrix. The accuracy metric was used to calculate the proportion of correct predictions relative to the total dataset, as shown in Equation (1).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Data} \quad (1)$$

Here, the number of correct predictions refers to classifications that match the actual labels previously assigned in the dataset. Accuracy provides a general overview of the model's ability to identify correct labels across different emotion classes. However, because data distribution in multi-class classification can be imbalanced, additional evaluation using the F1-score and confusion matrix was carried out to better capture classification quality in terms of precision and recall for each emotional category.

In addition to accuracy, this study also employed the Categorical Cross Entropy loss function as an indicator of how far the model's predictions deviate from the actual label distribution. This function measures the difference between the predicted probability distribution and the actual distribution using a logarithmic approach, as formulated in Equation (2).

$$Loss = -\sum_{i=1}^{N} y_i \,.\, \log(\hat{y}_i) \quad (2)$$

Where yi represents the actual label in one-hot encoded format, and $\hat{y}i$ is the predicted probability for that class. The resulting loss value reflects the extent of the model's error in predicting the correct class, with a lower loss indicating better model performance. This function was also employed during the training process as part of parameter optimization through backpropagation. By combining accuracy evaluation with the loss function, the model performance analysis becomes more comprehensive and in-depth, encompassing both the precision and robustness of the model in handling complex emotional data. A comparison of performance between baseline models CNN only and LSTM only and the combined CNN-LSTM model is presented in Table 4.

Table 4. Baseline Model vs CNN-LSTM Accuracy

Model	Accuracy (%)	F1-Score
CNN Only	72.4	0.70
LSTM Only	68.1	0.66
CNN-LSTM	85.7	0.84

IV. RESULT

A. Illustration

This section presents the performance evaluation results of the proposed CNN-LSTM hybrid model for affective gesture recognition within a VR environment. The evaluation was conducted by comparing this model against two baseline models, namely CNN-only and LSTM-only. The comparison was carried out quantitatively using standard classification evaluation metrics, such as accuracy, precision, recall, and F1-score. These metrics were selected to comprehensively assess the model's performance from various perspectives, including both accuracy and consistency in classification. In addition, visualizations in the form of a confusion matrix and emotion class distribution are included to provide a more comprehensive depiction of the model's performance. These visual tools serve to clarify classification patterns and potential errors among emotional classes, thereby strengthening the interpretation of the quantitative outcomes derived from the experiments.

The emotion recognition performance of the CNN-LSTM model demonstrates superior results compared to the baseline models. As shown in Table 5, the hybrid model consistently achieved the highest scores across all evaluation metrics, including accuracy, precision, recall, and F1-score. These findings indicate that the combined architectural approach is more effective in managing the complexity of the data. The comparison suggests that the integration of CNN, as a spatial feature extractor, with LSTM, as a temporal sequence processor, enhances the model's capability to identify intricate and emotionally expressive gesture patterns. This advantage highlights the synergy between CNN's ability to capture visual structure and LSTM's strength in understanding motion sequences and dynamics. The combination enables the model to generate richer and more contextualized feature representations, which are highly relevant in gesture-based emotion recognition tasks.

Table 5. Emotion Class Distribution

Emotion	Number of Samples	Percentage (%)
Happy	120	24%
Sad	100	20%

Angry	90	18%
Neutral	110	22%
Surprised	80	16%
Total	500	100%

The data distribution presented in Table 5 indicates that the number of samples across the emotional classes is relatively balanced. This condition is crucial to ensure fair and unbiased model training across all categories. A balanced dataset facilitates the model's ability to recognize each emotional category proportionally, without developing a tendency toward the majority class. Such proportional distribution allows for more equitable learning across all emotion types, thereby minimizing the risk of class imbalance, which could compromise the model's prediction accuracy. Moreover, a balanced dataset supports a more representative performance evaluation, as the results are not skewed by dominant performance in only one or two classes. In this context, data balance is a critical factor in achieving optimal and reliable classification outcomes.

Before presenting further analytical results, the comparative performance of the three models—CNN-only, LSTM-only, and the combined CNN-LSTM is visualized to provide a clearer depiction of the effectiveness of each approach. This visualization is designed to illustrate how well each model performs in identifying emotions based on users' gestures within a virtual environment. Figure 1 presents four commonly used evaluation metrics in classification tasks: accuracy, precision, recall, and F1-score. Each of these metrics reflects a different dimension of evaluation, ranging from overall prediction accuracy to the balance between correctly identified positive cases and the number of errors. This visualization offers a strong empirical foundation for assessing the extent to which each model can effectively fulfill its role in real-world scenarios. Through these graphical representations, readers can gain a more detailed understanding of the strengths and limitations of each model approach in the context of gesture-based emotion recognition.

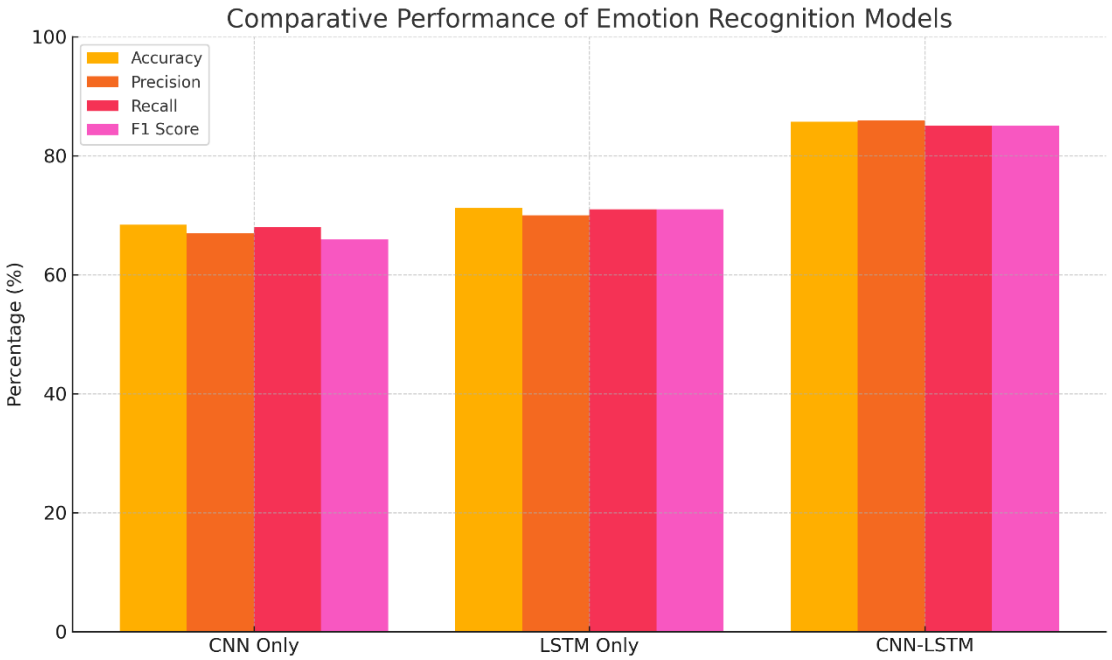

Figure 1. Comparative Performance of Emotion Recognition Models

As illustrated in Figure 1, the CNN-LSTM model consistently records the highest scores across all evaluated metrics. This indicates a significant performance improvement when spatial and temporal processing are integrated within a single model. The hybrid approach provides an advantage by simultaneously capturing the visual features of gestures and accounting for the temporal sequence of movements. The substantial performance gap between the CNN-LSTM and the other two models suggests that leveraging two complementary architectures can enhance classification quality. While CNN functions as a static feature extractor for each motion frame, LSTM simultaneously identifies temporal patterns occurring across frames. This combination demonstrates the model's effectiveness in understanding the dynamics of affective expressions compared to approaches that rely on only a single type of network architecture.

To gain deeper insight into how the model classifies each emotional category, an analysis was conducted using a confusion matrix. This matrix provides a visual representation of classification outcomes, mapping the number of correct and incorrect predictions for each class. The use of a confusion matrix allows for the identification of misclassifications between emotionally similar or overlapping categories. This representation is particularly useful for analyzing error patterns that may not be readily apparent from numerical metrics alone. Additionally, the matrix helps assess the model's ability to maintain classification consistency across different emotional states. The confusion matrix for the CNN-LSTM model is shown in Figure 2, offering a concrete illustration of how each emotion is classified, as well as the distribution of errors among the available categories. Through this approach, the model's performance can be understood more

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 Arpil 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 21-40

holistically, with a focus on real-world classification scenarios involving diverse affective expressions.

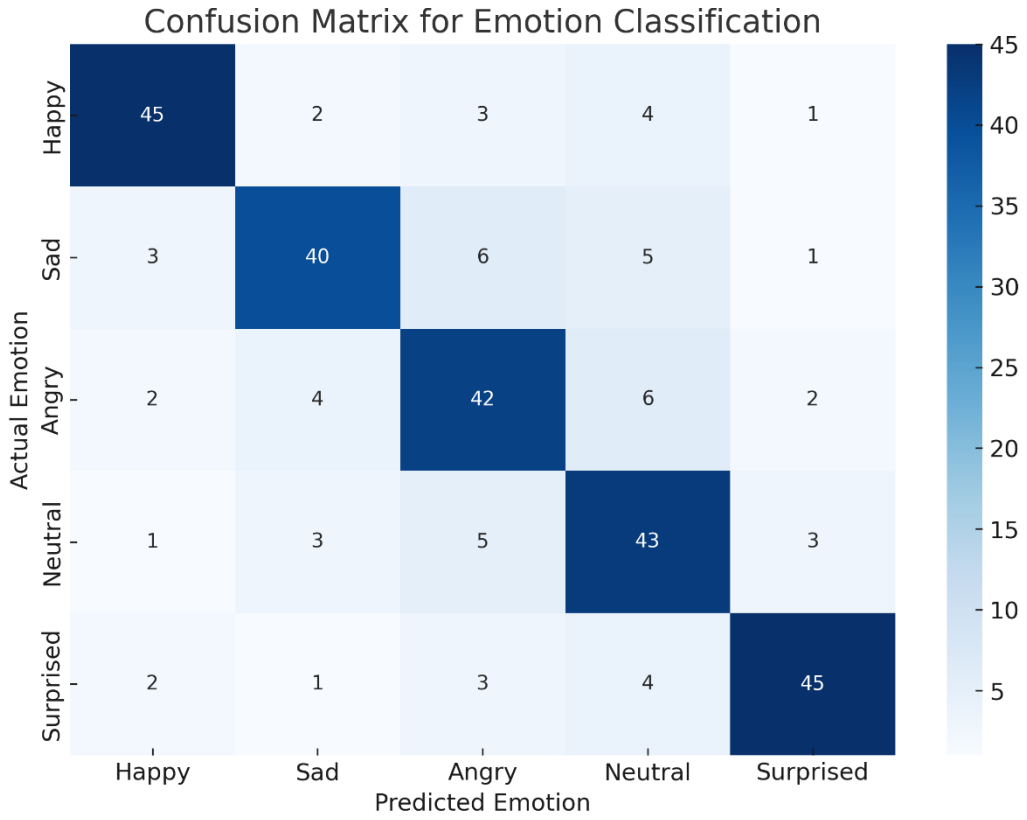

Figure 2. Confusion Matrix for Emotion Classification

The results presented in Figure 2 indicate that the model exhibits high accuracy in recognizing the happy and neutral emotions, as most predictions align along the main diagonal of the matrix. This pattern suggests that the model tends to classify these two emotions more accurately than the others. However, several misclassification cases were observed, particularly between the angry and sad classes, indicating that the gestural expressions of these emotions tend to share similarities in movement patterns. These errors highlight the challenge of distinguishing between emotions that may appear expressively similar in a gestural context. This finding reveals that the model's capacity remains limited when dealing with expressions that are visually and temporally proximate. The confusion matrix analysis thus provides insight into specific areas that require further improvement to enable the model to capture subtle affective expressions with greater accuracy.

The outcomes of this experiment demonstrate that the CNN-LSTM approach yields a significant improvement in the accuracy and precision of gesture-based emotion classification. The model not only performs well in identifying emotional states but also shows strong adaptability to the

diverse range of gestural expressions encountered in dynamic virtual environments. This capability reflects the effectiveness of combining spatial feature extraction through CNN with temporal sequence processing via LSTM in recognizing complex emotional patterns. Furthermore, the model's stable performance across various emotional categories indicates its ability to maintain classification quality despite variations in users' expressive behaviors. The balanced distribution of emotional data used in training also serves as a crucial supporting factor in ensuring more representative and unbiased outcomes. This condition enables the model to learn the distinctive characteristics of each emotion more equitably, contributing to more consistent classification performance across all observed emotional categories.

V. DISCUSSION

The findings of this study indicate that the combined CNN and LSTM model holds significant potential for recognizing emotion based on body gestures in VR environments, primarily due to its capability to process spatial and temporal data simultaneously. This reinforces the findings of (Huang et al., 2023), who reported that a similar approach achieved high accuracy in emotional gesture classification. In contrast to the SVM-based approach used by (Swoboda et al., 2022), deep learning models have proven to be more effective in capturing temporal dynamics that are crucial for gesture recognition. The success of the proposed model also lends support to the critiques put forward by (Kaklauskas et al., 2022) and (Atmaja et al., 2022), who noted that much of the research in affective computing remains focused on facial expressions and vocal cues, while body gestures remain underutilized. Gesture-based recognition systems such as the one proposed here offer new opportunities to enhance responsiveness in simulation-based training and VR therapy applications, aligning with the findings of (Yuvaraj et al., 2025), who emphasized the importance of emotional adaptation in digital interaction contexts. Consequently, the approach adopted in this study contributes to expanding the discourse on the development of more inclusive affective computing systems that accommodate non-verbal modalities.

The implementation of this system also demonstrates that gesture-based emotion processing can serve as an alternative or complement to conventional approaches that rely on facial expressions or voice analysis. For instance, the results of this study complement those of (Izountar et al., 2022) and (Dirin & Laine, 2023), which focused on facial and vocal emotion detection systems but did not incorporate users' kinesthetic expressions. On the other hand, these results also point to the potential for further integration between gesture recognition and other multimodal data to improve the accuracy and sensitivity of the system. More nuanced and adaptive emotional responses can be achieved by combining multiple sensory input channels within VR environments. This approach aligns with current trends in HCI research, which increasingly emphasize the

importance of natural and intuitive interaction, as discussed by (Strazdas et al., 2022), who highlighted gestures as a communication medium that closely resembles human interaction. This study thus broadens the application scope of affective computing and affirms the pivotal role of body gestures in constructing more empathetic and personalized user experiences in VR-based systems.

VI. CONCLUSION AND RECOMMENDATION

This study demonstrates that the CNN-LSTM model is highly effective in recognizing emotion based on users' body gestures in VR environments. The model's ability to capture emotional expressions through bodily movement facilitates the creation of more natural and intuitive interactions between users and VR systems. This is particularly important considering that verbal cues or facial expressions are often difficult to capture accurately in many VR contexts, making body gestures a primary source of emotional information. The integration of emotion recognition through body movement has the potential to enhance system responsiveness to users' emotional states, thereby improving comfort and engagement during VR use. The combination of CNN for spatial feature extraction and LSTM networks for temporal processing enables the system to handle complex and emotionally meaningful motion patterns. This potential paves the way for the development of affective computing systems that are not only technically intelligent but also emotionally perceptive in virtual environments.

For future development, applying this system in real-world scenarios, such as virtual therapy, simulation-based military training, or educational VR games, represents an important step in testing its reliability in dynamic settings. Real-time performance evaluation can provide critical insights into the model's adaptability and responsiveness to variations in user behavior. Additionally, improving the quality of the dataset, both in terms of sample size and emotional label diversity, will enhance the model's ability to recognize a broader spectrum of emotions. Incorporating physiological data such as heart rate and EEG signals could offer an additional dimension that strengthens the validity of multimodal emotion recognition. Exploring transfer learning techniques is also highly relevant for optimizing the training process, particularly when data is limited or computational efficiency is required. The use of attention-based models may further enhance the system's ability to focus on emotionally salient segments of gestures, thereby increasing the contextual accuracy of emotion recognition. Future studies are encouraged to compare various deep learning approaches to gain a deeper understanding of the most suitable methods for emotion recognition in complex and interactive virtual environments.

.

REFERENCES

Alabdullah, B. I., Ansar, H., Mudawi, N. Al, Alazeb, A., Alshahrani, A., Alotaibi, S. S., & Jalal, A. (2023). Smart Home Automation-Based Hand Gesture Recognition Using Feature Fusion and Recurrent Neural Network. *Sensors*, *23*(17), 7523. https://doi.org/10.3390/s23177523

Arman, A., Prasetya, P., Arifany, F. N., Pradnyaparamita, F. B., & Laksito, J. (2022). A DIGITAL PRINTING APPLICATION AS AN EXPRESSION IDENTIFICATION SYSTEM. *Journal of Technology Informatics and Engineering*, *1*(2), 5–15. https://doi.org/10.51903/JTIE.V1I2.135

Atmaja, B. T., Sasou, A., & Akagi, M. (2022). Survey on Bimodal Speech Emotion Recognition from Acoustic and Linguistic Information Fusion. *Speech Communication*, *140*, 11–28. https://doi.org/10.1016/j.specom.2022.03.002

Chouhayebi, H., Mahraz, M. A., Riffi, J., Tairi, H., & Alioua, N. (2024). Human Emotion Recognition Based on Spatio-Temporal Facial Features Using HOG-HOF and VGG-LSTM. *Computers*, *13*(4), 101. https://doi.org/10.3390/computers13040101

Dirin, A., & Laine, T. H. (2023). The Influence of Virtual Character Design on Emotional Engagement in Immersive Virtual Reality: The Case of Feelings of Being. *Electronics*, *12*(10), 2321. https://doi.org/10.3390/electronics12102321

Grewal, D., Herhausen, D., Ludwig, S., & Villarroel Ordenes, F. (2022). The Future of Digital Communication Research: Considering Dynamics and Multimodality. *Journal of Retailing*, *98*(2), 224–240. https://doi.org/10.1016/j.jretai.2021.01.007

Huang, Z., Ma, Y., Wang, R., Li, W., & Dai, Y. (2023). A Model for EEG-Based Emotion Recognition: CNN-Bi-LSTM with Attention Mechanism. *Electronics*, *12*(14), 3188. https://doi.org/10.3390/electronics12143188

Izountar, Y., Benbelkacem, S., Otmane, S., Khababa, A., Masmoudi, M., & Zenati, N. (2022). VR-PEER: A Personalized Exer-Game Platform Based on Emotion Recognition. *Electronics*, *11*(3), 1–16. https://doi.org/10.3390/electronics11030455

Kaklauskas, A., Abraham, A., Ubarte, I., Kliukas, R., Luksaite, V., Binkyte-Veliene, A., Vetloviene, I., & Kaklauskiene, L. (2022). A Review of AI Cloud and Edge Sensors, Methods, and Applications for the Recognition of Emotional, Affective and Physiological States. *Sensors*, *22*(20), 7824. https://doi.org/10.3390/s22207824

Kaseris, M., Kostavelis, I., & Malassiotis, S. (2024). A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition. *Machine Learning and Knowledge Extraction*, *6*(2), 842–876. https://doi.org/10.3390/make6020040

Khan, U. A., Xu, Q., Liu, Y., Lagstedt, A., Alamäki, A., & Kauttonen, J. (2024). Exploring Contactless Techniques in Multimodal Emotion Recognition: Insights into Diverse Applications, Challenges, Solutions, and Prospects. In *Multimedia Systems* (Vol. 30, Issue 3). Springer Berlin Heidelberg. https://doi.org/10.1007/s00530-024-01302-2

Kopalidis, T., Solachidis, V., Vretos, N., & Daras, P. (2024). Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. *Information*, *15*(3), 1356. https://doi.org/10.3390/info15030135

Leong, S. C., Tang, Y. M., Lai, C. H., & Lee, C. K. M. (2023). Facial Expression and Body Gesture Emotion Recognition: A Systematic Review on the Use of Visual Data in Affective Computing. *Computer Science Review*, *48*, 100545. https://doi.org/10.1016/j.cosrev.2023.100545

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. *Entropy*, *25*(10), 1440. https://doi.org/10.3390/e25101440

Mourtzis, D., Angelopoulos, J., & Panopoulos, N. (2023). The Future of the Human–Machine Interface (HMI) in Society 5.0. *Future Internet*, *15*(5), 162. https://doi.org/10.3390/fi15050162

Rahman, M. M., Gupta, D., Bhatt, S., Shokouhmand, S., & Faezipour, M. (2024). A Comprehensive Review of Machine Learning Approaches for Anomaly Detection in Smart Homes: Experimental Analysis and Future Directions. *Future Internet*, *16*(4), 139. https://doi.org/10.3390/fi16040139

Rani, C. J., & Devarakonda, N. (2022). An Effectual Classical Dance Pose Estimation and Classification System Employing Convolution Neural Network –Long ShortTerm Memory (CNN-LSTM) Network for Video Sequences. *Microprocessors and Microsystems*, *95*, 104651. https://doi.org/10.1016/j.micpro.2022.104651

Shomoye, M., & Zhao, R. (2024). Automated Emotion Recognition of Students in Virtual Reality Classrooms. *Computers & Education: X Reality*, *5*, 100082. https://doi.org/10.1016/j.cexr.2024.100082

Siddiqui, M. F. H., Dhakal, P., Yang, X., & Javaid, A. Y. (2022). A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database. *Multimodal Technologies and Interaction*, *6*(6), 47. https://doi.org/10.3390/mti6060047

Strazdas, D., Hintz, J., Khalifa, A., Abdelrahman, A. A., Hempel, T., & Al-Hamadi, A. (2022). Robot System Assistant (RoSA): Towards Intuitive Multi-Modal and Multi-Device Human-Robot Interaction. *Sensors*, *22*(3), 923. https://doi.org/10.3390/s22030923

Swoboda, D., Boasen, J., Léger, P. M., Pourchon, R., & Sénécal, S. (2022). Comparing the Effectiveness of Speech and Physiological Features in Explaining Emotional Responses during Voice User Interface Interactions. *Applied Sciences*, *12*(3), 1269. https://doi.org/10.3390/app12031269

Udahemuka, G., Djouani, K., & Kurien, A. M. (2024). Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. *Applied Sciences*, *14*(17), 8071. https://doi.org/10.3390/app14178071

Vrskova, R., Kamencay, P., Hudec, R., & Sykora, P. (2023). A New Deep-Learning Method for Human Activity Recognition. *Sensors*, *23*(5), 2816. https://doi.org/10.3390/s23052816

Yaseen, Kwon, O. J., Kim, J., Jamil, S., Lee, J., & Ullah, F. (2024). Next-Gen Dynamic Hand Gesture Recognition: MediaPipe, Inception-v3 and LSTM-Based Enhanced Deep Learning Model. *Electronics*, *13*(16), 3233. https://doi.org/10.3390/electronics13163233

Yuvaraj, R., Mittal, R., Prince, A. A., & Huang, J. S. (2025). Affective Computing for Learning in Education: A Systematic Review and Bibliometric Analysis. *Education Sciences*, *15*(1), 65. https://doi.org/10.3390/educsci15010065

Zheng, Y., & Blasch, E. (2023). Facial Micro-Expression Recognition Enhanced by Score Fusion and a Hybrid Model from Convolutional LSTM and Vision Transformer. *Sensors*, *23*(12), 5650. https://doi.org/10.3390/s23125650