Iraqi Journal of Computers, Communications, Control and Systems Engineering
Journal homepage: https://ijccce.uotechnology.edu.iq

Emotion Recognition from Upper-Body Movements with a Pose Collage CNN Architecture

Gheed Tawfeeq Waleed[1,*], Shaymaa Hameed Shaker[2]

[1,2]College of Computer Science, University of Technology, Baghdad, Iraq.
[1]cs.21.03@grad.uotechnology.edu.iq, [2]shaimaa.h.shaker@uotechnology.edu.iq

DOI: https://doi.org/10.33103/uot.ijccce.25.2.5

HIGHLIGHTS

- **Highlights**
- Proposed a collage-based Convolutional Neural Network (CNN) framework that transforms temporal motion into spatial representations using keyframe selection, enabling efficient and accurate upper-body emotion recognition.
- Achieved a maximum classification accuracy of **95.72%** on the BoLD dataset, outperforming existing body-based emotion recognition methods.
- Demonstrated the reliability of upper-body movement as an independent modality for emotion detection, especially in scenarios where facial or vocal cues are unavailable or ambiguous.
- Highlighted promising applications in human–computer interaction, behavioral analysis, security, and assistive technologies.

ARTICLE HISTORT

Received: 23/ April /2025

Revised: 30/ May /2025

Accepted: 05/ July /2025

Available online: 30/ October /2025

Keywords:
Body Movement, Convolution Neural Network, Deep learning, Emotion Recognition, Skelton.

ABSTRACT

Emotion recognition through body movement presents a compelling alternative to traditional facial and vocal analysis in the field of emotion recognition . This study introduces a deep learning-based framework that utilizes upper-body movements features to classify emotional states using Convolutional Neural Networks (CNNs). The system is trained and evaluated on the Body Language Dataset (BoLD) dataset, targeting seven primary emotions: happiness, sadness, anger, fear, surprise, joy, and disgust. The proposed model achieves a classification accuracy of 95.72%, significantly outperforming existing methods in body-based emotion recognition. The result highlights the potential of body posture as a reliable and standalone modality for emotion detection, especially in contexts where facial information is ambiguous or unavailable. The presented work contributes to the advancement of non-intrusive, vision-based affective computing systems, with promising applications in human-computer interaction, behavioral analysis, security, and assistive technologies. Future research may explore multimodal integration, real-time deployment, and cross-cultural adaptability to further enhance the robustness and versatility of body-driven emotion recognition systems.

I. INTRODUCTION

The ability to recognize and interpret human emotions is fundamental to natural interactions, whether between individuals or between humans and machines. While facial expressions and speech-based emotion recognition have been widely studied, research has shown that body movements also play a crucial role in conveying affective states [1]. In many real-world environments—including faces that are occluded or masked, poor lighting conditions, or limited audio input—these sources may be unreliable or even entirely absent. Studies suggest that nonverbal cues, including gestures, posture, and movement patterns, contribute significantly to emotional communication, sometimes even more than facial expressions alone. Mehrabian's widely accepted communication model suggests that 55% of communication is nonverbal, with body language being a key component. However, despite its importance, emotion recognition through body movement remains relatively underexplored compared to facial and vocal expression analysis.[2]

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have led to breakthroughs in visual-based emotion recognition.[3] While prior research has primarily focused on facial expression datasets such as FER2013 [4] and AffectNet[5], body movement-based emotion recognition is gaining traction with datasets like FABO[6] , GEMEP[7], and BoLD[8]. The BoLD dataset, in particular, provides a comprehensive collection of body language expressions captured in natural settings, making it an ideal resource for training deep learning models to recognize emotions through upper-body movements. However, existing methods face challenges in generalizability, due to the complexity of body motion patterns and the lack of standardized annotation protocols for body-driven emotions.

The body can convey emotional intent through signs of affective information in posture, gesture, and movement dynamics; however, it has not yet fully established itself as an independent channel of communication. Despite its promise, the existing body-based emotion recognition researches are constrained by limitations in both scaled datasets, and accuracy. Many existing systems rely on hand-crafted features or are trained on limited, constrained datasets, resulting in reduced generalizability. This identifies a notable research gap: the absence of efficient and scalable deep learning models that can autonomously learn emotional representations from body motion data [10]. Addressing the issue is essential for developing robust multimodal emotion recognition systems capable of functioning reliably in uncontrolled real-world environments.Expanding the focus to include body movement as a core modality can lead to more adaptive, context-aware, and inclusive affective computing solutions. This research introduces a CNN-based deep learning framework to classify emotions using upper-body movement analysis.

Key Contributions of The proposed Study:

- Development of a robust CNN model optimized for body movement-based emotion recognition, utilizing keyframe extraction and feature selection to enhance performance.

- Comparison with existing state-of-the-art approaches, demonstrating superior accuracy in classifying emotions through upper-body movements.

- Exploration of real-world applications, including human-computer interaction (HCI), behavioral analysis, and automated surveillance, where nonverbal emotion cues play a critical role.

The structure of the paper is as follows: The text that follows Section II will cover the research conducted on related work.
Section III: outlines the proposed CNN-based framework, covering data preprocessing, model architecture, and training methodology. Section IV describes the experimental outcomes.
Section V: discusses findings and compares the model's performance with previous approaches.
Section VI: concludes the paper and suggests directions for future research.

II. RELATED WORK

Emotion recognition through body movements has gained significant attention due to its potential applications in human-computer interaction, mental health diagnostics, and social robotics. Recognizing body movements provides further contextual and affective content information compared to facial expressions and vocal analysis, enhancing recognition accuracy when facial cues are absent or ambiguous. With the newly developed machine learning and motion capture techniques, the automatic recognition of emotions from body movements and poses has gained momentum. Nonetheless, issues persist, such as data limitations, cultural variability, and real-time applicability. Wu et al. [11] Investigated various approaches to enhance emotion recognition from body movements by training emotion classification machine learning models using Laban Movement Analysis (LMA) to differentiate emotions based on human movement. Their dataset captured movement dynamics like acceleration, trajectory, and posture shifts, identifying them as the important elements for differentiating happiness, sadness, anger, and fear. But the researchers found certain subtle emotions — disappointment and contentment, for instance — shared common movement characteristics, making it difficult to distinguish between them. Luo et al. in another study Body Language Dataset (BoLD) (2019), over 9000 video clips recorded in natural environments depict human body expressions. They analyzed movement patterns using deep learning models (CNNs and RNNs), resulting in an average recognition accuracy of 72%. Nonetheless, their study focused on dataset bias since the majority of samples were exaggerated emotions, which could have limited the model's generalization to subtle and spontaneous expressions [12].

Piana et al. introduced a real-time skeletal-based classifier utilising 3D joint data, achieving an accuracy of 61.3%. However, its computational overhead limited its application in real-time scenarios. [13]. Additionally Ahmed et al. [14] introduced a hybrid model that integrates ANOVA and a genetic algorithm for the purpose of feature selection. Despite achieving over 90% accuracy for walking sequences, 96.0% for sitting postures, and 86.66% for an activity-independent analysis, generalisation across different environments remained a limitation. . The study demonstrated effective, real-time hand movement classification, but showed challenges with maintaining accuracy in multiple movement contexts. BEE-NET [15] developed a multi-modal deep learning model that integrates facial, body, and contextual evidence. The method, while powerful, is highly dependent on facial shapes and consequently lacks robustness in occluded conditions.

Chitra et al. [16] presented a real-time deep learning architecture utilising LSTM and pose estimation to interpret sign language, demonstrating the model's effectiveness in analysing sequential body pose features. Their method serves as a valuable reference for time-series processing in affective recognition.
Tsai et al.[17] introduced a novel graph convolutional network framework, the Spatial Temporal Variation Graph Convolution Network (STV-GCN), which enhances traditional ST-GCN models by capturing variations in skeleton movement speed over time. The recognition of fine-grained emotional behaviours, such as fear and sadness, was achieved through the identification of optimal motion tempos.

This was accomplished by training a convolutional neural network that enables the learning and modulation of convolution filters at each frequency, resulting in significantly higher accuracy.

Despite these advancements, body-based emotion recognition has some problems that prevent it from reaching its full potential. Dataset diversity is another are out of the box, since most data are collected in controlled environments, with actors mimicking predetermined gestures. The results in the absence of spontaneous everyday data that better approximates natural emotional expressions. One of the reasons is that cultural differences strongly influence body language interpretations to the extent that it is not easy to create a universal model. Moreover, some gestures are ambiguous, as the same movement may have different meanings depending on social context [18]. Our methodology utilises a collage-based CNN design and converts temporal motion into spatial representation via keyframe selection and entropy filtering, thereby enhancing efficiency while achieving state-of-the-art accuracy. Table I represents a summary of key studies on body based emotion recognition , including method used, dataset and key limitations .

TABLE I. A COMPARATIVE OVERVIEW OF BODY-BASED EMOTION RECOGNITION METHODS

Study	Year	Dataset	Method	Accuracy	limitations
Luo et al.[12]	2019	BoLD	CNN+RNN	72%	• Biased dataset • Extreme emotions
Wu et al. [11]	2020	LMA dataset	Speech Analysis + machine learning	70 %	Weak discrimination among subtle emotions
Piana et al. [13].	2014	3D Skeletal	Geometric Features + Classifier	61.3 %	High computation
Ahmed et al. [14]	2019	Recorded dataset	ANOVA + Genetic Algorithm	90%	Poor generalization
Dehshibi et al [15]	2024	Multi source	CNN	66.3 %	Depends on facial features
Chitra et al [16]	2025	Custome	LSTM	89.3 %	Requires streaming input; evaluated using sign language
Tsai et al [17]	2021	Skeleton	STV-GCN	75%	requires tempo-labeled data; computationally intensive for extensive datasets

III. METHODOLOGY

The proposed method utilises upper body movement to identify human emotions by extracting and processing pose features via a Convolutional Neural Network (CNN). This paper provides a comprehensive overview of the dataset, preprocessing pipeline, pose estimation procedure, key-frame selection strategy, input transformation into collage images, the adopted CNN architecture, and the training setup utilised for model optimisation. *Fig. 1* presents a comprehensive overview of the methodology, detailing the sequential process of converting raw video footage into emotion classification.

A. Dataset And Data Acquisition

In this work, BoLD dataset was analyzed, which contains a total of 9,876 video clips aggregated over 27 different emotional states. Each clip captures spontaneous human behavior in naturalistic environments, offering a rich source of real-world data. For the purpose of the research, seven primary

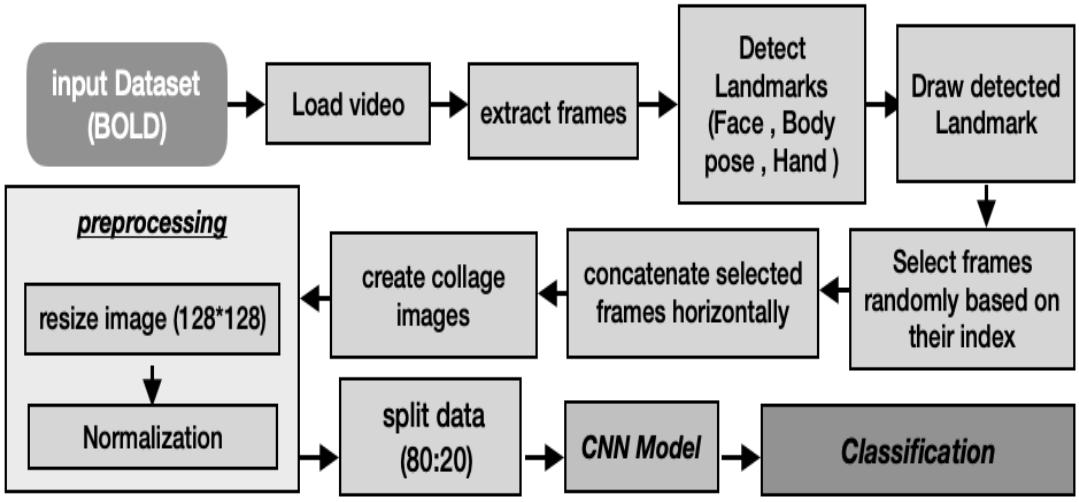

FIG. 1. PROPOSED ARCHITECTURE OF THE RECOGNITION SYSTEM.

emotions—*happiness, sadness, anger, fear, surprise, joy,* and *disgust*—were selected based on Ekman's

basic emotion model [19] . The dataset provides sufficient temporal information to support detailed analysis of motion patterns associated with emotional expression, making it highly suitable for training and evaluating deep learning models in body-based emotion recognition.Table II represents a summary of Bold dataset.

TABLE II. EMOTION DISTRIBUTION DATA OF BOLD

Emotion category	No. of Clips	Total frames
Happy	1250	5625
Sad	1100	4620
Anger	1000	4200
Fear	900	3900
Surprise	950	4275
Natural	1050	4515
Disgust	950	4095
Total	7200	31230

B. Pose Estimation And Landmark Detection

To extract spatial representations of upper-body movement, the study employed a pose estimation pipeline using MediaPipe Holistic, a real-time framework capable of jointly tracking 468 facial landmarks, 33 body pose landmarks, and 21 keypoints per hand across video frames. MediaPipe achieves high landmark localization accuracy by leveraging lightweight neural networks for real-time inference, even under challenging environmental conditions.

Each frame from the BoLD video clips was processed through MediaPipe to generate a high-resolution skeletal map of the upper body. The key landmarks used in the study included:

- Face: Eyes, eyebrows, nose, mouth corners, jawline
- Upper Body: Head, neck, shoulders, elbows, wrists, torso
- Hands: Fingertips, palm center, and wrist joints

These extracted landmarks were visualized and overlaid on the original frames to produce annotated keypoint images, as illustrated in *Fig. 2.* These overlays preserve the temporal and spatial context of body motion while removing background noise and irrelevant visual data.

In total, 522 landmarks were tracked per frame, but only those corresponding to upper-body and hand regions were retained for further processing. This selective focus ensures the model concentrates on movement dynamics most indicative of emotional expression, while also reducing computational complexity. Frames were processed at a resolution of 256×256 pixels, and the resulting landmark coordinates were stored for both visualization and downstream feature encoding.To enhance temporal consistency and minimize the impact of frame jitter or motion blur, a Savitzky-Golay filter was applied to smooth the trajectories of joint positions across consecutive frames. The smoothing process effectively reduces the influence of outliers and better captures meaningful motion patterns associated with different emotional states.

FIG. 2. EXAMPLE OF FACIAL AND UPPER-BODY LANDMARKS EXTRACTED USING MEDIAPIPE.

C. Keyframe Selection And Collage Generatio

The application of CNN to each frame of every video sequence was both computationally intensive and resulted in redundancy among temporally consecutive frames. Therefore, an entropy-based selection method was employed to construct a representative emotional expression from a limited number of moments.[20] This methodology reduces the temporal footprint while preserving variance in movement and emotional significance.

Instead of sampling frames evenly distributed in each video, frames that have maximum motion variability or pose uncertainty were extracted to represent the whole emotional expression space. The frames map onto distinct phases of motion: onset, peak, transition, and resolution. The fixed-interval sampling removes dependence on emotion-specific templates or heuristics as cut-off values, which makes it universally applicable across emotion types and subjects. Shannon entropy is calculated for each frame, in order to determine information content and thus select keyframes as demonstrated in equation (1) :

$$H(f) = -\sum_{i=1}^{N} pi \ log_2 pi \quad (1)$$

H(f): Entropy of frame, pi: Probability of pixel intensity ,N: How many discrete levels of intensity there are in the frame.

The chosen frames are annotated with pose landmarks and then horizontally concatenated to create a single collage image. There are two major advantages to the design:

It captures the temporal evolution of motion while maintaining the spatial relationships in each frame. It enables the CNN to learn spatiotemporal dependencies without having to rely on recurrent structures (e.g. LSTMs, GRUs) or 3D convolutions (complex).
Each collage image is 128 × 640 in size (stitched five 128×128 annotated frames horizontally). This compact representation successfully turns the temporal classification problem into a spatial one, which naturally fits standard CNN-based architectures.

Pixel values were normalized as well as optional data augmentation methods such as horizontal flipping, slight rotation, and brightness shifting were applied before passing collages to the network. These augmentations improve the model's generalization ability and mitigate overfitting at the training stage.

D. Pre-Processing Data

Deep learning models require the input data to be clear, uniform, and appropriate, which can be handled through preprocessing. For this study, a few preprocessing operations were applied to the collage images generated from the selected keyframes. All these are done to improve training efficiency, enforce generalization, and fit the neural architecture to the CNN structure.

Preprocessing plays a crucial role in preparing input data that is clean, standardized, and compatible with deep learning models. There are several preprocessing operations that were applied to the collage images generated from selected keyframes. These steps were designed to enhance training efficiency, promote generalization, and ensure architectural compatibility with the CNN.

***i.* Image Resizing**

The size of each collage image (obtained by horizontally concatenating five pose-annotated keyframes) was then resized to a fixed size of 128 × 128 pixels. The downsampling of the feature maps is done by bilinear interpolation to reduce the spatial resolution while preserving edge information, and this idea allows faster training and lower memory consumption . Standardizing the input size ensured uniformity across training samples and compatibility with the CNN's input layer.

***ii.* Normalization**

To improve training stability and convergence speed, pixel values were normalized to the range [0, 1] by dividing each pixel by 255 as shown in equation 2 . the normalization process reduces input variance, helping activation functions—particularly Rectified Linear Unit ReLU—operate within a numerically stable range. It also mitigates inconsistencies caused by lighting variation and differing brightness levels across the BoLD dataset.

$$x_{norm} = \frac{x}{255} \tag{2}$$

X norm : normalized pixel value, x :original pixel value.

***iii.* Data Augmentation**

To reduce overfitting and improve the model's generalization ability, data augmentation was applied to the training set. The following techniques were used:

- Horizontal flipping with a probability of 0.5 to simulate mirrored gestures
- Random rotation within ±10 degrees to account for camera tilt or slight postural shifts
- Brightness jittering to simulate varying lighting conditions
- Zoom scaling between 90% and 110% to reflect motion scale variability

These transformations were applied on-the-fly during training using Keras' ImageDataGenerator, ensuring that each epoch received a unique, dynamically augmented batch of training samples.

***iv.* Dataset Partitioning**

To evaluate the effect of training data size on performance, multiple training/testing splits were tested. The primary configuration used an 80% training / 20% testing ratio, which yielded the best results in preliminary trials. Additional configurations, including 70:30 and 75:25, were explored for comparison.

Splits were shuffled and stratified to keep a similar distribution of emotion classes across train and test sets. Such stratification helped mitigate any risk of label imbalance, which can skew the assessment of a model's performance.

- Convolutional Neural Network Architecture And Training Process
- Model Architecture

The architectural design of a Convolutional Neural Network (CNN) is a critical component in optimizing its performance for complex tasks such as emotion recognition. Designing this involves determining the number of convolutional layers, pooling layers, fully connected layers, and activation functions, as well as their interconnections. Convolutional Neural Networks (CNNs) represent a category of feedforward neural networks that employ convolutional operations to autonomously acquire hierarchical feature representations from input data. Artificial neurones in CNNs, modelled after biological neurones, utilise kernels (or filters) as receptive fields to analyse local patterns in spatial dimensions. Activation functions, such as ReLU, serve as thresholds that determine whether to activate a neurone by simulating signal transmission while optimising performance based on weights acquired during training.[20] CNNs are specifically designed for visual or spatiotemporal tasks, as the integration of these layers forms a computational pipeline that facilitates the extraction, condensation, and interpretation of spatial features. This section outlines the CNN architecture and training pipeline, detailing the procedures employed to optimise model performance, including preprocessing, parameter tuning, loss function selection, and regularisation methods. The collaged inputs enable the model to comprehend motion through a constrained yet diverse collection of images, facilitating a potential compression of the intrinsic data represented by the moving components, thereby serving as an efficient and accurate framework for emotion integration.

The study employs a CNN architecture that initiates with three consecutive convolutional layers utilising 3×3 kernels and the ReLU activation function to extract pertinent spatial features related to body posture and motion patterns, ranging from low to high levels : The first convolutional layer uses 128 filters, The second layer uses 64 filters, The third layer uses 32 filters where the convolution feature map output value is computed as shown in equation 3 :

$$f(i,j) = \sum_{m=1}^{M} \sum_{n=1}^{N} x(i+m, j+n) \cdot k\,(m,n) \qquad (3)$$

x(i+ $m, j+n$) *: Input pixel value at the kernel's offset , k(m,n) : filter value. , F(I,j) : output feature map value at position (i,j), M*N : kernel matrix size .*

A 2×2max-pooling layer follows each convolutional layer, lowering spatial resolution and highlighting dominant features it is applied as illustrated in equation 4 :

$$p(i,j) = max_{\,(m,n)\epsilon R}\, f(i+m, j+n) \qquad (4)$$

P(i,j) : Output after pooling , R : pooling window , f(i+m, j+n) : Feature values of input within region R.

To prevent overfitting, dropout layers are used after each pooling stage with rates of 0.45 after the first pooling layer, 0.35 after the second, 0.25 after the third layer

After the last convolutional block, the flattened layer takes the 3D feature maps and flattens them into a 1D vector used as input for two fully connected layers:

- The first dense layer contains 128 neurons
- The second dense layer has 32 neurons Both use ReLU activation, and batch normalization is applied after each dense layer to stabilize learning and speed up convergence.

The final classification is handled by a softmax output layer with seven neurons, corresponding to the target emotion categories: happiness, sadness, anger, fear, surprise, joy, and disgust. The complete structure of the proposed Convolutional Neural Network (CNN) is illustrated in *Fig. 3*, which depicts the sequential arrangement of key components used for emotion classification based on upper-body movement inputs.

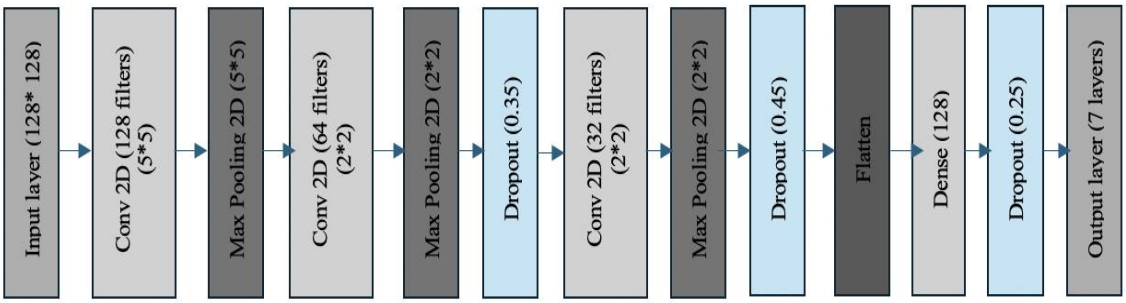

FIG. 3. OVERVIEW OF THE PROPOSED CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE.

- Model Training Process

The CNN model is trained using supervised learning to minimise the distance between predicted emotion labels and actual emotion labels. The categorical cross-entropy loss function is utilised, as it is suitable for multi-class classification tasks such as emotion recognition.

The Adaptive Moment Estimation (ADAM)[21], an optimization algorithm that computes adaptive learning rates for each parameter using both AdaGrad and RMSProp. It's used popularly due to its effectiveness and low memory cost when training deep neural networks. A minimal learning rate of 0.0001 is employed to ensure stable weight updates and prevent gradient exceeding during the initial training phases.

Weights are initialised with the Xavier uniform initialiser, which ensures a balance in the variance of incoming and outgoing activations in each layer. The approach maintains a smooth gradient flow, thereby promoting effective training. which allows for flexibility in training duration and convergence characteristics, with the model being trained over various epochs using batch sizes of 32 and 64. The dataset is partitioned using various training-to-testing ratios, such as 70:30 and 80:20, to assess the model's generalisation capability. The training set comprises data utilised iteratively to adjust model parameters, while the test set serves to independently assess performance on data that remains unknown to the model.

Dropout Regularisation [22] is employed to mitigate overfitting by randomly deactivating neurones during training, thereby discouraging the model from excessively depending on specific pathways within the feature space. Batch normalisation is utilised to stabilise learning and mitigate internal covariate shift, thereby enhancing the convergence of training.

Although no early stopping mechanism is enforced, the training process is manually monitored across epochs to detect signs of underfitting or overfitting, allowing for adjustments as needed.

The training strategy is part of an end-to-end structured pipeline that includes : Raw video loading, Frame extraction, Pose estimation, Collage generation, Data normalization

The resulting processed collage images are then fed into the CNN model for training and inference. The pipeline is designed to ensure consistent training across diverse body expressions and to promote robust generalization to unseen emotional movements.

The dimensions of the convolution kernel size in the first convolutional layer were 128 and then 64 and 32 in the last convolution layer .it should be smaller than those of the input image. During the convolution process, a feature map is generated by applying each kernel to the input image. In essence, the convolution kernel traverses the input image, computing the dot product at each spatial position. The feature maps generated by each kernel are combined in the depth dimension to form the output image of the convolutional layer. The proposed model uses convolutional layers with dropout (0.45, 0.35, and 0.25) also batch normalization (64 and 32) to improve performance and generalization, regularization to avoid overfitting, and a final softmax layer for classification. Finally, it examines the complete input matrix and provides a return. Max pooling systematically decreases the spatial dimensions of the representation (output features matrix) and facilitates the extraction of dominant features, resulting in spatial invariance. Incorporating essential body expression patterns into the model enhances its ability to classify emotions more accurately. The final step in classification involves assigning the input data to one of seven distinct classes: sadness, joy, surprise, fear, anger, disgust, or happiness.

IV. RESULT AND DISCUSSION

This section evaluates the proposed CNN-based emotion recognition model empirically using the BoLD dataset. In order to understand the impact of various training configurations on the model's classification performance, including the number of epochs, batch sizes, and the ratios of data split. All the experiments were performed on a personal computer with an Intel® Core™ i7-8550U CPU @ 1.80 GHz, 16 GB RAM under 64-bit Windows operating system. It was trained and assessed over three training durations, 80, 100, and 120 epochs respectively. For each duration of the training, various proportions was tested between the data splits according to 70:30, 80:20, 75:25 as the training and testing data proportions. For each of the tests, training accuracy, training loss, test accuracy and validation loss were logged. The arrangement grants insights on the model's performance across different training conditions, allowing us to track the model's behavior in different configurations that best suit its robustness for inferring emotion from upper-body movements.

A. Experimental Setup And Training Evaluation

Table III summarizes the hyperparameters and how it affercts the accuracy rate and loss for the bold database to human emotion recognition. The data was split 70% for training and 30% for testing, using a batch size of 32. An experiment was performed over three epochs—80, 100 and 120—where the impact of train length on model performance was examined. The range of training accuracy during this process was 95.4% to 97.3%, meaning that learning continued across the epochs.

Test accuracy varied between 91.4% and 95.3%, reflecting strong generalization to unseen data. The validation loss values ranged from 0.1853 to 0.4591, showing a clear reduction in error with longer training cycles, which suggests improved model convergence and stability over time. The analysis involves seven categories,

TABLE III. RESULTS UNDER 70:30 TRAIN-TEST SPLIT

Data split (70:30)	Epoch	Train accuracy	Train loss	Test accuracy	Validation loss
	80	0.9732	0.1022	0.9359	0.4591
	100	0.9677	0.0914	0.9530	0.1853
	120	0.9543	0.0796	0.9145	0.3446

As shown in Table III, the model achieved its best performance at 100 epochs under the 70:30 split, demonstrating a strong alignment between training and test accuracy. This consistency indicates that the model effectively learned generalizable features from the training data without overfitting, resulting in reliable performance on unseen test samples. Furthermore, the loss ratio was analyzed and concluded that the test loss is 0.1853.

In the second configuration (Table IV), the training and testing split was adjusted to 80:20, and the batch size was increased to 64. The setup produced the highest test accuracy among all configurations evaluated. When trained for 120 epochs, the model achieved a training accuracy of 98.1% with a minimal training loss of 0.0563, indicating strong learning efficiency. Correspondingly, the test accuracy peaked at 95.72%, demonstrating exceptional generalization to unseen data. The validation loss also remained relatively low, further supporting the model's stability and robustness under this configuration.

TABLE IV. RESULTS UNDER 80:20 TRAIN-TEST SPLIT

Data split (80:20)	Epoch	TRAIN ACCURACY	TRAIN LOSS	TEST ACCURACY	VALIDATION LOSS
	80	0.9492	0.1452	0.9415	0.2478
	100	0.9747	0.0844	0.9415	0.1006
	120	0.9810	0.0563	0.9572	0.2064

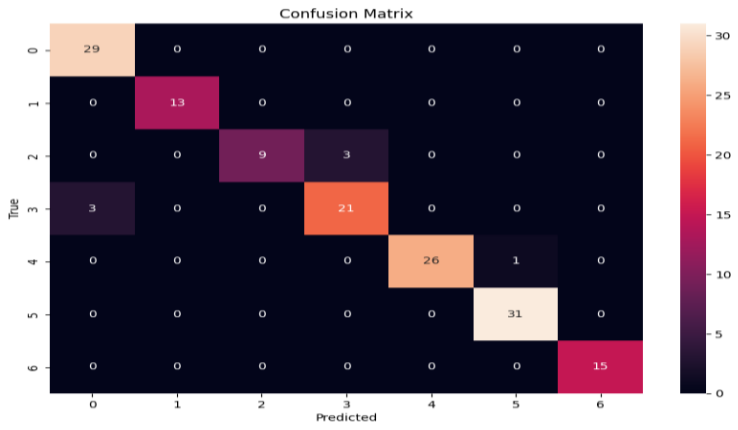

FIG. 4. CONFUSION MATRIX OF THE PROPOSED MODEL (70:30).

The third configuration (Table V) used a 75:25 split. The model exhibited consistently strong performance, reaching training accuracy levels up to 98.2% and test accuracy of 97.4% at 100 epochs. While validation loss varied across training durations, the overall trend showed reduced error as training progressed.

TABLE V. RESULTS UNDER 70:30 TRAIN-TEST SPLIT

Data split (80:30)	Epoch	TRAIN ACCURACY	TRAIN LOSS	TEST ACCURACY	VALIDATION LOSS
	80	0.9449	0.1592	0.9017	0.5160
	100	0.9752	0.0899	0.9344	0.3405
	120	0.9827	0.0656	0.9530	0.2870

The confusion matrices generated for each experimental configuration, as shown in *Fig. 4, 5, and 6,* provide a detailed view of the class-wise performance of the proposed CNN model under different train-test split ratios. Each matrix visualizes the distribution of predicted versus actual emotion labels, where darker diagonal cells indicate a higher number of correctly classified instances, and off-diagonal cells represent misclassifications.

Across all configurations, the model exhibited a strong diagonal dominance, indicating high precision and recall for most of the seven emotion classes. However, the degree of performance varied depending on the data split ratio and number of epochs used during training.

Under the 70:30 split (*Fig. 4*), the model achieved solid overall performance, but slight confusion was detected between ‘fear’ and ‘surprise’, as well as ‘joy’ and ‘happiness’. This confusion likely arises from the overlap of upper-body gestures common to these emotions, such as raised shoulders, expanded chest, or sudden movements, which are characteristic of both positive and reactive emotions.

In the case of the 80:20 split (*Fig. 5*), the matrix displayed a sharper diagonal with fewer off-diagonal elements, indicating superior generalization capabilities. The outcome is consistent with the observed higher test accuracy (95.72%) and lower validation loss under this configuration. The misclassification rates were significantly reduced, especially for challenging emotion pairs like ‘fear’ vs. ‘surprise’.

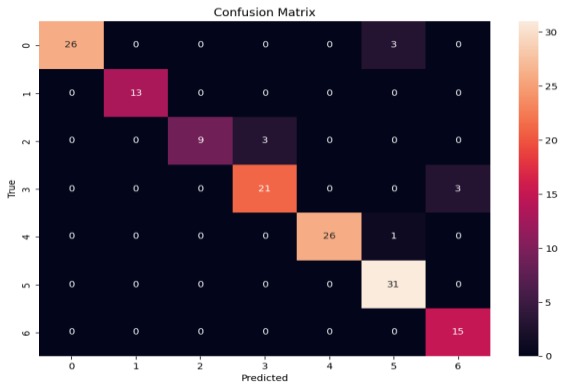

FIG. 5. CONFUSION MATRIX OF THE PROPOSED MODEL (80:20).

For the 75:25 split (*Fig. 6*), the model maintained strong classification capabilities, although a slight decrease in performance compared to the 80:20 configuration was observed. Most misclassifications involved emotions with subtle body language differences, particularly ‘anger’ and ‘disgust’, which often share similar postural indicators, such as forward-leaning and upper-body tension.

A notable trend across all matrices is the model's consistent ability to accurately classify emotions like ‘sadness’ and ‘happiness’. These emotions typically involve distinctive and less ambiguous postural cues—such as slouched shoulders for sadness and an open chest posture for happiness—which the model effectively learned to recognize.

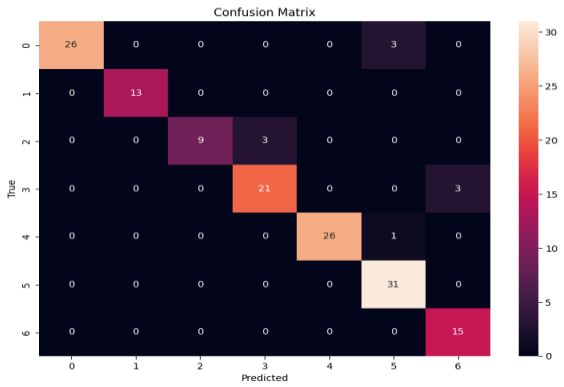

FIG. 6. CONFUSION MATRIX OF THE PROPOSED MODEL (75:25).

The comparative analysis confirms that the proposed CNN system not only achieves strong performance in terms of aggregate metrics (accuracy and loss) but also demonstrates robust class-level discrimination. Furthermore, the observed confusion patterns suggest that future research could benefit from integrating additional temporal or dynamic features, such as the velocity or acceleration of keypoints, to further enhance the model’s ability to distinguish between emotionally similar classes.

B. Discussion And Comparssion

This section presents a comprehensive analysis of the results obtained, accompanied by a comparison of the proposed system with existing studies in the field of emotion recognition from body movement. Section IV illustrated that the model's performance surpassed the benchmarks of previous methods on the BoLD dataset, highlighting the approach's effectiveness.

i. Model Interpretation and Insights

The proposed CNN-based system achieved consistently high accuracy across all experimental configurations, with the best performance occurring under the 80:20 split, where the model attained a test accuracy of 95.72% at 120 epochs. The improvement can be attributed to three core factors: (1) the extraction of high-resolution pose landmarks using MediaPipe, (2) the collage-based representation of keyframes that captures temporal motion patterns in a single static input, and (3) the use of regularization techniques, including dropout and batch normalization, which helped mitigate overfitting and improve generalization.

The model exhibited stable learning dynamics, with both training and validation losses decreasing steadily as training progressed. Notably, the 80:20 split provided a favorable balance between the amount of training data and the diversity of the test set, likely contributing to the superior generalization compared to the 70:30 and 75:25 configurations.

ii. Confusion Matrix Analysis

The confusion matrices (*Fig. 4–6*) displayed a strong diagonal dominance, confirming that most emotion classes were accurately classified. However, minor misclassifications were consistently observed between ‘fear’ and ‘surprise’, and occasionally between ‘joy’ and ‘happiness’. These errors

can be attributed to similarities in upper-body movements across these emotions, such as rapid arm gestures, raised shoulders, or open-arm postures, which tend to overlap in expressive patterns.
Despite these minor confusions, the model showed high precision for emotions like ‘sadness’ and ‘happiness’, which are often associated with more distinctive postural cues, making them easier for the model to separate.

iii. Comparative Evaluation

To objectively assess the model's performance, Table VI presents a comparison with baseline and state-of-the-art methods on the BoLD dataset. While previous models such as DCNN, BEE-NET, and ST-GCN achieved moderate results, they struggled when limited to upper-body motion cues. In contrast, the proposed collage-based CNN framework significantly outperformed these methods without relying on explicit temporal modeling techniques like RNNs or GCNs.

V. CONCLUSIONS

The study presented a deep learning-based framework for emotion recognition using upper-body movements, leveraging convolutional neural networks (CNNs) to classify emotional states. The model was trained and evaluated on the BoLD dataset, with a specific focus on seven primary emotions: *happiness, sadness, anger, fear, surprise, joy,* and *disgust*. By extracting pose-based motion features from video frames and transforming them into structured collage inputs, the system effectively learned discriminative patterns linked to each emotional category.

Through systematic experimentation with varying training configurations, the model achieved a maximum classification accuracy of 95.72%, underscoring the capability of CNNs to learn rich body posture representations. These results confirm the viability of body movement as a reliable and independent modality for affective state recognition, particularly in contexts where facial expressions are occluded, ambiguous, or absent.

The findings extend the scope of emotion recognition beyond traditional facial and vocal cues, highlighting applications in healthcare monitoring, human-computer interaction, behavioral assessment, biometric security, and robotics. Future research will focus on the generalisation of full-body and audio facial cues, the incorporation of adaptive keyframe selection, and the validation of datasets across different cultures and real-world scenarios.

CONFLICT OF INTEREST

The authors declare no conflict of interest in relation to this paper.

TABLE. VI. COMPARATIVE PERFORMANCE OF BODY-BASED EMOTION RECOGNITION METHODS ON BOLD DATASET

Method	Dataset	Reported Accuracy	Reference
DCNN	BoLD	76.80%	[12]
BEE-NET	BoLD	66.33%	[15]
ST-GCN	BoLD	68.20%	[12]
Baseline CNN	BoLD	65.00%	[12]
CNN	BOLD	97.2%	[23]
Proposed Method	BoLD	95.72%	-

LIST OF ABBERVATIONS

Abbervation	Full Form
CNN	Convolutional Neural Networks
BoLD	Body Language Dataset
FABO	Facial and Body expression database
GEMEP	Geneva multimodal Emotion Portrayals
LMA	Laban Movement Analysis
RNN	Recurrent Neural Networks
ANOVA	Analysis of Variance
BEE-NET	Bodily Expression Emotion Network
LSTM	Long short term memory
ST-GCN	Spatio- Temporal Graph Convolutional Network
HCI	Human Computer Interaction
ReLU	Rectified Linear Unit
ADAM	Adaptive Moment Estimation
CPU	Central Processing Unit
RAM	Random Access Memory

REFERENCES

[1] I. Pikoulis, P. P. Filntisis, and P. Maragos, "Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild," in *Proc. 16th IEEE Int. Conf. Automatic Face and Gesture Recognition (FG)*, 2021, doi: 10.1109/FG52635.2021.9666957.

[2] Y. Zeng, N. Xiao, K. Wang, and H. Yuan, "Real-time facial expression recognition using deep convolutional neural network," in *Proc. IEEE Int. Conf. Mechatronics and Automation (ICMA)*, 2019, doi: 10.1109/ICMA.2019.8816322.

[3] J. Z. Wang *et al.*, "Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion," *Proc. IEEE*, vol. 111, no. 10, 2023, doi: 10.1109/JPROC.2023.3273517.

[4] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Information Processing (ICONIP)*, Daegu, Korea, 2013, pp. 117–124.

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.

[6] H. Gunes and M. Piccardi, "Bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Proc. Int. Conf. Pattern Recognition (ICPR)*, 2006, doi: 10.1109/ICPR.2006.39.

[7] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: The multimodal emotion recognition test (MERT)," *Emotion*, vol. 9, no. 5, pp. 691–704, 2009, doi: 10.1037/a0017088.

[8] J. Dhamala *et al.*, "BOLD: Dataset and metrics for measuring biases in open-ended language generation," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2021, doi: 10.1145/3442188.3445924.

[9] M. A. Ali, A. J. Hussain, and A. T. Sadiq, "Human body posture recognition approaches," *ARO–Sci. J. Koya Univ.*, vol. 10, no. 1, 2022, doi: 10.14500/aro.10930.

[10] F. Noroozi *et al.*, "Survey on emotional body gesture recognition," *IEEE Trans. Affective Comput.*, vol. 12, no. 2, pp. 505–523, 2021, doi: 10.1109/TAFFC.2018.2874986.

[11] C. Wu, D. Davaasuren, T. Shafir, R. Tsachor, and J. Z. Wang, "Bodily expressed emotion understanding through integrating Laban movement analysis," *Patterns*, vol. 4, no. 10, 2023, doi: 10.1016/j.patter.2023.100816.

[12] Y. Luo *et al.*, "ARBEE: Towards automated recognition of bodily expression of emotion in the wild," *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 1–25, 2020, doi: 10.1007/s11263-019-01215-y.

[13] S. Piana, A. Staglianò, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," Feb. 2014. [Online]. Available: arXiv:1402.3270.

[14] F. Ahmed, A. S. M. H. Bari, and M. L. Gavrilova, "Emotion recognition from body movement," *IEEE Access*, vol. 8, pp. 11761–11771, 2020, doi: 10.1109/ACCESS.2019.2963113.

[15] M. M. Dehshibi and D. Masip, "BEE-NET: A deep neural network to identify in-the-wild bodily expression of emotions," *arXiv preprint*, arXiv:2402.12345, Feb. 2024.

[16] P. Chitra, P. Brindha, K. Srilatha, G. Jegan, and R. V. Raju, "LSTM based pose estimation and sign language translation," in *Proc. Int. Conf. Trends in Material Science and Inventive Materials (ICTMIM)*, 2025, pp. 1830–1835.

[17] M. F. Tsai and C. H. Chen, "Spatial temporal variation graph convolutional networks (STV-GCN) for skeleton-based emotional action recognition," *IEEE Access*, vol. 9, pp. 13 870–13 877, Jan. 2021.

[18] S. Alwajidi and L. Yang, "Multi-resolution hierarchical structure for efficient data aggregation and mining of big data," in *Proc. Int. Conf. Automation, Computational and Technology Management (ICACTM)*, 2019, doi: 10.1109/ICACTM.2019.8776717.

[19] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971, doi: 10.1037/h0030377.

[20] A. W. Majeed, S. H. Shaker, and A. A. Saeid, "A real time face recognition and tracking framework using lightweight convolutional neural network," *BIO Web Conf.*, vol. 97, p. 00029, Apr. 2024, doi: 10.1051/bioconf/20249700029.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, Dec. 2014.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] V. K. Singh, J. Barman, S. Kumar, and J. Jayadeva, "CoRE-BOLD: Cross-domain robust and equitable ensemble for BOLD signal analysis," in *Proc. Mach. Learn. Health (ML4H)*, 2025, pp. 961–975.