Surveys in Mathematics and its Applications

ISSN 1842-6298 (electronic), 1843-7265 (print)
Volume 21 (2026), 79 – 105

A MULTIMODAL DEEP LEARNING FRAMEWORK FOR HUMAN BEHAVIOR RECOGNITION AND SYNTHESIS USING CNN-LSTM AND ENSEMBLE MODELS

Vaikunta Pai T, Manjula Mallya M, Ramona Birau, Nethravathi P S, Virgil Popescu, Iuliana Carmen Bărbăcioru, Pramod Vishnu Naik

Abstract. This study integrates deep learning models to represent, analyze, and generate diverse human behaviors, including postures, gestures, facial expressions, physiological signals, and emotional states. By modeling multimodal signals, the research develops a holistic framework for understanding and recreating complex human behaviors, advancing human-computer interaction (HCI) and enabling empathetic, responsive digital experiences. This approach offers transformative applications across healthcare, education, entertainment, security, automotive, and human resources. In healthcare, it supports patient well-being monitoring, while in education, it enables personalized learning experiences. Entertainment benefits from the creation of immersive, emotionally resonant content, and security sectors gain improved threat detection capabilities. In the automotive field, this research can inform advanced driver-assistance systems (ADAS), enhancing vehicle safety, while in human resources, it supports improved team dynamics and productivity. By prioritizing multimodal data integration, the study enhances accuracy in behavior recognition and the efficient processing of large-scale data. These advancements not only elevate HCI by making interactions more natural and intuitive but also support the development of tailored, human-centered applications. This work paves the way for a future where technology authentically replicates the depth of human expression, fostering an empathetic, adaptive digital environment that responds meaningfully to individual needs.

1 Introduction

Understanding human behavior is a foundational aspect of social interaction, communication, and intelligent decision-making. Human body language-comprising pos-

2020 Mathematics Subject Classification: 68T07; 68T37; 68T50
Keywords: Convolutional Neural Networks (CNNs), Human Behavioral Data, Long Short-Term Memory (LSTM), Deep Learning Models, Multimodal signals, Human Body Language, human-centric applications

tures, movements, gestures, facial expressions, vocal tones, physiological signals, and emotional states-offers critical insights into individuals' intentions, mental states, and interpersonal dynamics. Accurately capturing and interpreting these multimodal behavioral cues is essential for the development of intelligent systems capable of empathetic, adaptive, and human-aware responses. The digitization of human activities has led to large volumes of behavioral data from sources such as videos, wearable sensors, social media, and ambient environments. Extracting meaningful patterns from this high-dimensional data remains a key challenge in affective computing and behavioral analysis.

In recent years, deep learning has emerged as a transformative approach in behavior modeling due to its ability to automatically learn hierarchical representations from large-scale data. Specifically, Convolutional Neural Networks (CNNs) have shown remarkable success in visual understanding tasks such as gesture recognition, facial emotion classification, and action recognition [34],[39]. Additionally, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been widely adopted for modeling temporal dynamics in human behavior, especially when analyzing video streams or sequential sensor data [34]. Recent advancements in attention mechanisms and transformer-based models have further improved the contextual understanding of complex, multimodal behavioral patterns [6] .

The integration of deep learning models into behavior understanding has shifted the paradigm from handcrafted features to end-to-end learning systems that combine multiple modalities. For instance, Zadeh et al. proposed a Multimodal Transformer for emotion recognition by jointly modeling language, vision, and acoustic signals [9]. Similarly, Li et al. introduced PoseC3D, which applies 3D convolutional models to skeletal data for effective human action recognition in diverse environments [14]. These approaches have opened up new frontiers for real-time human-computer interaction, adaptive learning systems, mental health diagnostics, and personalized digital experiences.

This study proposes a framework to model and extract multimodal signals from human behavioral data. It also enables data-driven synthesis of lifelike behaviors, such as postures, gestures, and emotional expressions. Thus, enabling intuitive and naturalistic human-computer interactions. By leveraging the strengths of CNNs, RNNs, attention-based mechanisms, and ensemble architectures, the proposed methodology explores the representation, recognition, and generation of human behavioral data with high fidelity and real-world applicability.

These contributions have wide-ranging implications. In healthcare, behavioral modeling supports early detection of mental health issues and assists in patient monitoring [5]. In education, it enables the creation of adaptive learning environments that respond to students' engagement levels [30]. In autonomous systems, behavior-aware AI enhances safety and interaction in fields like robotics and automotive design [24]. As such, this work contributes to the growing field of human-centric AI, where machines understand, anticipate, and support human needs through naturalistic in-

teraction paradigms.

Research Questions. Delving into this multifaceted exploration of human behavioral data and its applications through deep learning raises several critical research questions.

RQ1: How can deep learning models effectively represent and encode multimodal signals in human behavioral data, including postures, movements, gestures, facial expressions, and physiological signals?

RQ2: How can deep learning models generate realistic and expressive human postures, gestures, and facial expressions based on a given context or emotion?

RQ3: How can the transformative potential of deep learning-driven human behavior analysis be harnessed to develop human-centric applications in healthcare, education, and entertainment? What are the specific challenges and opportunities in each of these domains?

RQ4: What ethical considerations should be taken into account when generating and collecting human behavioral data, particularly when employing deep learning models? How can privacy and security concerns be addressed effectively?

These research questions underscore the complexity and significance of our journey. They serve as guiding beacons, illuminating the path toward a future where technology not only recognizes but authentically replicates the richness of human expression. In doing so, we aim to enhance human-computer interaction, shaping a digital world that is more empathetic, responsive, and meaningful while addressing ethical and practical considerations in this endeavor.

The representation, analysis, and generation of human behavioral data are the focal points of this research. Specifically, our exploration centers on the utilization of deep learning models, which have demonstrated remarkable potential in decoding and predicting human behavior across a spectrum of domains. From healthcare to social networks, and even water resources systems, these models have harnessed multimodal signals from human body language, including postures, movements, gestures, facial expressions, sounds, physiological signals, and emotional states.

This research aims to explore four primary aspects related to the representation, analysis, and generation of human behavioral data:

1. Modeling and Extraction of Multimodal Signals in Body Language:

- Investigating deep learning architectures to effectively extract and represent multimodal signals from various sources of human behavioral data;
- Exploring techniques to fuse different modalities, such as visual, auditory, and physiological signals, to capture a comprehensive understanding of human behavior.

2. Recognition and Understanding of Human Body Language:

- Expanding the current state-of-the-art in deep learning algorithms for the recognition and understanding of human body language;
- Developing models capable of accurately recognizing and interpreting different postures, movements, gestures, and facial expressions.

3. Data-Driven Synthesis for Posture, Gesture, and Expressions:

- Exploring generative deep learning models to synthesize realistic and expressive human postures, gestures, and facial expressions;
- Investigating techniques for generating high-fidelity behavioral data, enabling the creation of diverse training datasets for deep learning models.

4. Human-Centric Applications in Healthcare, Education, and Entertainment:

- Examining the potential applications of human behavioral data analysis in healthcare, such as patient monitoring, emotion detection, or rehabilitation;
- Investigating educational applications where deep learning models could assist in assessing student engagement, attention, or understanding;
- Exploring the integration of deep learning-generated human behavioral data in entertainment industry applications like virtual reality, animation, or gaming.

By conducting this research, the aim is to contribute to the advancement of human body language understanding, leveraging the power of deep learning models. The outcomes of this study will provide valuable insights for developing intelligent systems that can interpret, generate, and utilize human behavioral data in various domains.

2 Related Work

The exploration of human behavioral data represents a multifaceted endeavor, encompassing a diverse spectrum of postures, movements, gestures, facial expressions, sounds, physiological signals, and emotional states. This research domain has transcended disciplinary boundaries, holding profound implications across various sectors, including healthcare, education, and entertainment. With the pervasive digitalization of society, the integration of deep learning models has catalyzed transformative advancements in representing, analyzing, and generating human behavioral data [39].

Past research has embarked on a comprehensive journey beyond mere data analysis, striving to construct holistic frameworks capable of modeling, extracting, and understanding multimodal signals inherent in human life. This includes ventures

into the nuanced domains of recognizing and comprehending human body language, as well as pioneering data-driven synthesis techniques to recreate lifelike postures, gestures, and expressions for facilitating human-computer interaction [12].

In recent years, significant strides have been made in leveraging deep learning techniques to analyze and understand human behavioral data. Notably, DeepMind's WaveNet has demonstrated the capacity of deep generative models to synthesize natural-sounding speech, facilitating advancements in speech recognition and synthesis [26]. Similarly, Microsoft's Seeing AI and Google's Duplex showcase the practical applications of deep learning in enhancing accessibility and enabling more natural human-computer interactions through intelligent conversational agents [27]. Emotient's facial expression recognition technology underscores the importance of deep learning in interpreting human emotions from facial cues, with implications for various domains, including healthcare and market research [24]. Additionally, MIT's DeepMimic project explores the use of deep reinforcement learning to simulate human-like movements and behaviors, offering promising avenues for applications in gaming, animation, and robotics [14]. These pioneering works collectively highlight the transformative potential of deep learning in deciphering and synthesizing human behavioral data, paving the way for more sophisticated and empathetic AI systems.

At the core of these methodologies lie deep learning architectures, renowned for their ability to discern intricate patterns within heterogeneous datasets. Convolutional Neural Networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms serve as the cornerstone of this research, enabling the fusion of diverse modalities within human behavioral data. Deep learning models have showcased remarkable potential in understanding and predicting human behavior across domains such as healthcare, social networks, and water resources systems [34],[37]. For instance, dual-stream 3D CNNs have shown enhanced accuracy in action recognition by capturing both spatial and motion-based features in human movement datasets [23].

To further refine behavior understanding, emotion recognition models have integrated CNNs with attention mechanisms to dynamically focus on expressive facial regions, improving classification accuracy in affective computing tasks [23]. Meanwhile, transformer-based architectures are being optimized for real-time behavioral inference on edge devices, enabling efficient deployment in latency-sensitive applications like smart surveillance and health monitoring [18].

This research endeavors to address critical research questions concerning the effective representation and encoding of multimodal signals in human behavioral data, the generation of realistic human postures, gestures, and facial expressions, and the harnessing of deep learning-driven human behavior analysis for developing human-centric applications in healthcare, education, and entertainment [9]. Furthermore, ethical considerations surrounding the generation and collection of human behavioral data, particularly when employing deep learning models, are paramount. Ensuring privacy, security, and mitigating biases are essential aspects of this endeavor [29], es-

pecially as real-world deployments demand ethical alignment with user expectations and regulatory standards [18].

However, previous research in human behavioral data extraction has encountered several challenges that have limited the effectiveness of traditional approaches. These challenges include issues related to feature representation, scalability, generalization, interpretability, and ethical considerations. Traditional methods often struggle to capture the complex and nuanced nature of human behavior, leading to limited accuracy and reliability in data extraction. Additionally, scalability becomes a concern when dealing with large datasets, hindering the applicability of these methods to real-world scenarios. Moreover, the interpretability of results obtained from traditional approaches is often lacking, making it difficult to understand the underlying patterns in the data. Ethical considerations, such as privacy and bias, also pose significant challenges in the context of human behavioral data extraction. However, the emergence of Deep Learning offers promising solutions to these challenges. Deep Learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated superior performance in feature learning, scalability, generalization, and interpretability. By leveraging the power of Deep Learning, this research aims to overcome the limitations of traditional approaches and advance the state-of-the-art in human behavioral data extraction.

Overall, this research navigates the complex terrain of human behavioral data, leveraging deep learning models to enhance our understanding of human expression and interaction. By addressing critical research questions and ethical considerations, it seeks to pave the way for a future where technology authentically replicates the richness of human expression, fostering empathetic, responsive, and meaningful human-computer interaction.

3 Methodology

The methodology employed in this paper draws upon the robust capabilities of Deep Learning Convolutional Neural Networks (CNNs) to extract human behavioral data from visual data sources, such as images and video recordings. CNNs, inspired by the organization of the animal visual cortex, have demonstrated exceptional performance in various computer vision tasks due to their ability to automatically learn hierarchical representations of data [21]. CNNs are composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers. In the context of human behavior analysis, these networks can be trained to recognize complex patterns and features within visual data, enabling the extraction of meaningful insights [7].

The methodology can be summarized into several key steps, each of which plays a crucial role in the successful extraction and analysis of behavioral data.

3.1 Data Collection

The initial step involves the collection of diverse and relevant visual data sources. This includes images and video footage obtained from various settings, such as controlled laboratory environments or real-world scenarios. The dataset comprises a wide spectrum of human behaviors, ensuring the model's robustness and versatility in capturing different actions and activities.

In the pursuit of understanding and analyzing human behavior using Deep Learning Convolutional Neural Networks (CNNs), the availability of comprehensive and well-annotated datasets is paramount [9]. Numerous public datasets have been curated and made accessible by various research groups, offering a diverse range of human behaviors captured in different contexts as shown in Table 1. These datasets serve as invaluable resources for training and evaluating CNN models for behavior recognition. For example, benchmark datasets such as UCF101, HMDB51, and NTU RGB+D 120 are widely used in action recognition and provide critical support for training deep networks on human motion [23]. More recent collections like the Kinetics-700 dataset further expand this scope, enabling training on diverse actions under real-world conditions [38].

Dataset	Description
Weizmann Dataset	Contains videos of various human actions like walking, running, and jumping, performed by a single individual in a controlled environment.
UCF101	Comprises video clips of diverse human actions such as sports, cooking, and dancing, making it suitable for action recognition using CNNs.
HMDB51	Consists of video clips of human actions across 51 categories, including actions like jogging, dancing, and martial arts, making it suitable for action recognition tasks with CNNs.
Charades Dataset	Includes video clips of people engaging in daily activities and communication, offering a rich source for studying human behavior and interactions using CNNs.
SBU Kinect Interaction Dataset	Provides depth and RGB data capturing social interactions between individuals, making it a valuable resource for analyzing social behavior using CNNs.

Table 1: Datasets for Human Behavioral Data Extraction.

In the research, experiments were conducted primarily based on three key datasets: the Weizmann dataset, the UCF101 dataset, and the HMDB51 dataset. The primary focus throughout this study revolved around the recognition of human behavior, with

a particular emphasis on the analysis of surveillance videos and images.

Weizmann Dataset: The Weizmann dataset encompasses ten distinct classes, each featuring nine videos captured using a static camera to analyze individual behaviors in everyday situations. In total, this dataset involves nine participants, and the image samples maintain a resolution of 180x144 pixels. To narrow our research scope, we considered some classes that represent various human behaviors, including walking, skipping, running, jumping jacks, Waving, and jumping. Figure 1 provides visual examples from the Weizmann dataset.

Figure 1: Sample Images from the Weizmann Dataset

UCF101 (Action Recognition Data Set) Dataset: The UCF101 dataset represents a significant resource in the field of action recognition, offering a diverse collection of 13,320 video clips capturing a wide spectrum of human actions and activities. This dataset is characterized by its extensive variety, encompassing 101 distinct action categories that span sports, physical exercises, dancing, and numerous other human movements. Recorded in real-world settings, UCF101 videos introduce realism into action recognition research, with challenges stemming from varying camera viewpoints, lighting conditions, and backgrounds. The dataset includes video clips of varying durations, providing a comprehensive range of action scenarios. Researchers often utilize UCF101 to train, validate, and test action recognition models, including those based on advanced deep learning architectures like Convolutional Neural Networks (CNNs). The dataset serves as a vital benchmark for evaluating the capability of action recognition algorithms to handle the complexities and diversities inherent in real-world action scenarios. Figure 2 provides visual examples from the UCF101dataset.

HMDB51 (Human Motion Database 51) Dataset: Similar to UCF101, the HMDB51 dataset consists of video clips capturing human actions across 51 distinct categories. These categories encompass various activities such as jogging, dancing, and martial arts. HMDB51 serves as an essential benchmark for research related to action recognition, providing a comprehensive dataset for the analysis of human behavior through CNN-based approaches. Figure 3 provides visual examples from the HMDB51dataset.

In the course of this research, the initiative was taken to curate own samples,

Figure 2: Sample Images from the UCF101Dataset

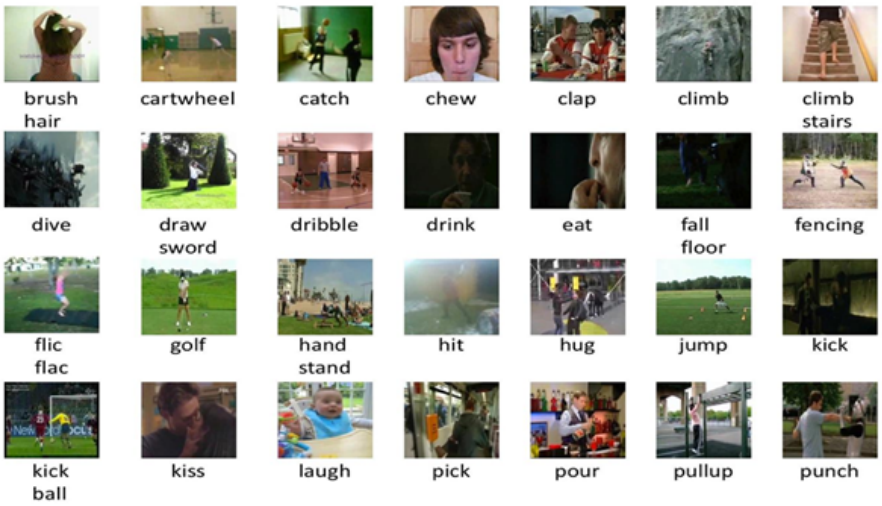

Figure 3: Sample Images from the HMDB51Dataset

driven by distinct motivations. Firstly, the creation of these custom datasets served as a critical step in rigorously testing the stability and robustness of our proposed models within real-world scenarios. By subjecting our models to the complexities of custom-captured data, we aimed to ensure their practical applicability and reliability. Furthermore, one of our key motivations for generating proprietary datasets was rooted in the limitations often encountered with publicly available datasets. Many existing public datasets are constrained by low-resolution video content, which may not accurately reflect the visual intricacies encountered in real-life scenarios. Custom datasets, in contrast, feature significantly higher resolutions, aligning more closely with contemporary standards and providing a more realistic and relevant basis for the research. These datasets were carefully selected based on their relevance to our research objectives in human behavior recognition. Utilizing these datasets, we developed and evaluated CNN models capable of recognizing and interpreting a wide array of human actions and behaviors within surveillance videos and images.

Figure 4 shows the flowchart of human behavior recognition by using deep learning. This flowchart presents a comprehensive overview of the process involved in deep learning-based human behavior recognition. It starts with the initial phase of Data Preparation, where video streams are transformed into frames, followed by crucial steps like Preprocessing, Region of Interest (ROI) Detection, Feature Extraction, Data Augmentation, and Data Splitting to prepare the dataset for model training.

In the Deep Learning Model section, the architecture is meticulously designed, and the model is trained, fine-tuned, and evaluated for optimal performance. The core of the process is Behavior Recognition and Classification, where the trained model is employed to classify human behaviors, followed by Temporal Analysis and Post-Processing to ensure accuracy and consistency in behavior predictions. Ultimately, the Output Results phase presents the recognized behaviors and classifications, making the deep learning-based human behavior recognition system a valuable tool for real-world applications and research.

3.2 Data Preparation

In the pursuit of extracting meaningful insights from human behavioral data through the utilization of Deep Learning Convolutional Neural Networks (CNNs), meticulous data preparation serves as a foundational phase. This phase encompasses a series of crucial steps, commencing with the extraction of video streams into individual frames. Subsequently, we delve into the realm of preprocessing, where frames undergo transformations to ensure consistency and optimal analysis conditions. The identification of Regions of Interest (ROIs) within frames, coupled with precise labeling, further refines our dataset for accurate recognition. Feature extraction is pivotal, as it encapsulates the essence of behavioral nuances. Data augmentation techniques are applied to enhance diversity and model robustness. To facilitate model training, we judiciously split the data into training, validation, and testing sets. This comprehensive data preparation process forms the bedrock upon which our Deep Learning CNN models will be built, enabling the extraction of nuanced human behavioral data from video streams with precision and efficacy.

In the pursuit of extracting human behavioral data through Deep Learning CNN, a systematic algorithmic approach is employed. Initially, the algorithm initializes a video reader to access the target video source. Subsequently, it calculates the total number of frames in the video, ensuring a comprehensive analysis. The second step is to manually find the region of interest (ROI) in each frame and label the ROI with correct class. We use a Python based toolbox for video frame labelling, which makes labelling much easier and time efficient. After finding the ROI, we create a rectangle that contains the correct class, and label each frame. Moreover, the name of each class is predefined by yourself ensure organized storage, the algorithm creates a designated directory if it doesn't already exist. The heart of the process lies in the meticulous processing of each frame within the video. Looping through each frame,

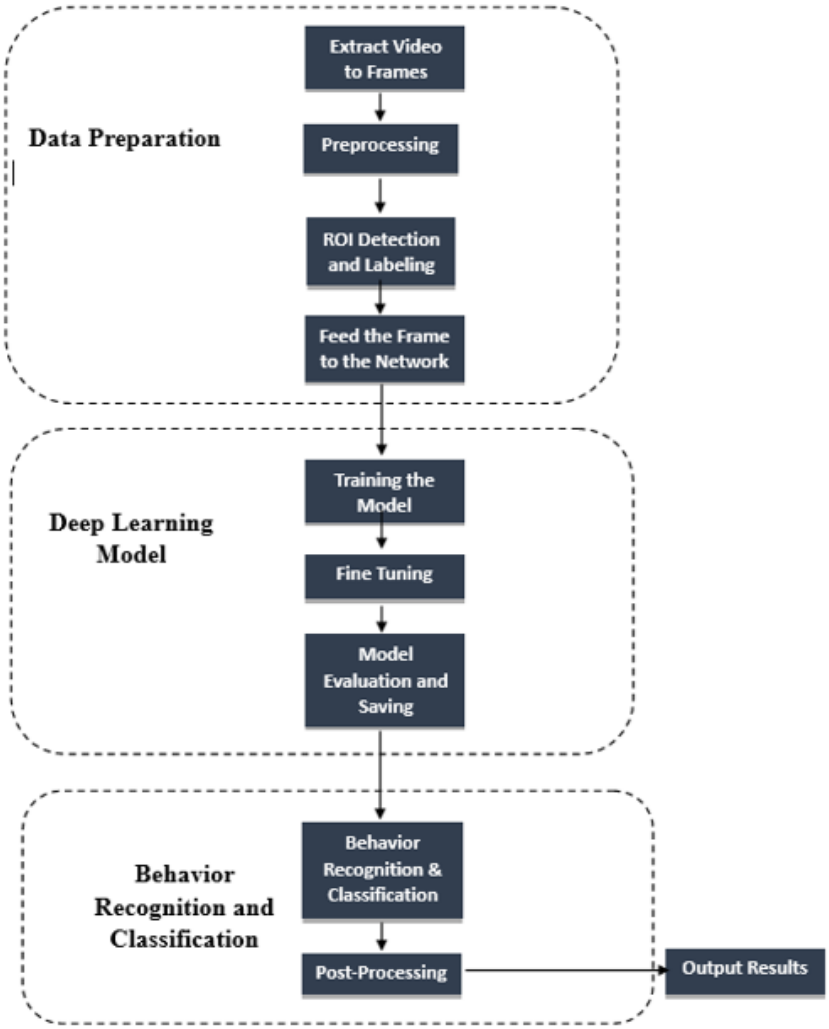

Figure 4: Deep Learning-Based Human Behavior Recognition Process

the algorithm captures and saves it with precision, assigning each frame a unique and informative name. This step, crucial to our research, forms the foundation for subsequent deep learning analysis. Error handling mechanisms are also incorporated to gracefully manage any unforeseen exceptions. Ultimately, this algorithm contributes significantly to the systematic extraction of human behavioral data, a cornerstone of our research endeavors.

Algorithm 1 Frame Extraction from Video

Require: video_file (path to input video), save_path (output directory)
Ensure: Frames saved to the specified directory

Initialize the video reader
video ← VideoReader(video_file)
Calculate total number of frames
frame_number ← ⌊ video.Duration × video.FrameRate ⌋
Create save directory if it does not exist
if not exist(save_path, 'dir') **then**
    mkdir(save_path)
end if
Process each frame
for $i = 1$ to frame_number **do**
    image_name ← strcat(save_path, num2str(i))
    image_name ← strcat(image_name, '.jpg')
    I ← read(video, i)
    imwrite(I, image_name, 'jpg')
    $I \leftarrow \emptyset$
end for
Finished
return

4 The Proposed Method

In the research, an innovative method is introduced for the extraction of human behavioral data using Convolutional Neural Networks (CNNs). This method represents a pivotal step towards a holistic understanding of human behavior, acknowledging its multi-modal nature encompassing text, image, and various data sources.

The approach integrates deep learning techniques, including CNNs, to capture intricate patterns within diverse datasets. By fusing information from different modalities, we aim to decipher the complex interplay of human actions and emotions. Furthermore, our method pioneers data-driven synthesis, facilitating the recreation of lifelike behavioral expressions. Real-time application is a core focus of our method, enabling its utilization in healthcare, education, entertainment, and beyond. This adaptability to real-world scenarios is a testament to the agility and responsiveness of the approach.

4.1 Convolutional Neural Network Model

Deep learning has a multi-level structure [29] that can handle complex feature extraction problems. As a typical model of deep learning, convolutional neural networks

have been widely used in many fields such as speech recognition, natural language processing, and pattern recognition [22],[11]. CNNs have also demonstrated superior performance in behavior recognition and real-time video analytics due to their ability to learn spatial hierarchies from pixel-level input [15]. The convolutional neural network is a multi-layered deep network, generally consisting of a convolutional layer and a pooled layer, and finally connected to the fully connected layer. This alternation between convolution and pooling layers enables CNNs to learn increasingly abstract representations while maintaining spatial relationships [5].

In the convolution process, after input data are fed into the network, they are processed by multiple trainable convolution kernels, resulting in feature maps that highlight important local patterns. The function of the convolutional layer is feature extraction. The neurons of each convolutional layer are connected with the data in the local receptive field of the previous layer to extract localized features. Each convolution kernel is trained to identify a specific type of visual characteristic, such as edges, corners, or texture. After the convolutional layer, the pooling layer performs a down-sampling operation on the resulting feature map. The purpose of down-sampling is to retain the most significant features while reducing the spatial dimensions, which leads to reduced computational cost and better generalization [35]. This special network structure allows the convolutional neural network to achieve high recognition rates across a wide range of applications.

The experiment includes the standard Neural Network as follows:

1. Text Input Branch (Optional): This branch processes textual data (It is included for the simplicity of the model) and it plays a crucial role in processing textual data, which is often a valuable source of behavioral insights. It takes text as input and converts it into a numerical format using techniques like tokenization and word embedding. In this example, we use an LSTM (Long Short-Term Memory) layer to model sequential text data.

2. Image Input Branch: This branch is designed for image data that is the frames extracted from the video stream, which provides rich visual information about human behavior. It employs convolutional layers to detect features like edges, shapes, and textures within images. Max-pooling layers are used to reduce the spatial dimensions of feature maps while preserving the most important information. This step helps the neural network focus on essential image features, making it effective in recognizing behavioral patterns from visual data. The combination of convolutional layers and max pooling is a standard approach for image feature extraction due to its effectiveness in capturing hierarchical features.

3. Concatenation (Optional): After processing text and image data separately, their features are concatenated or combined into a single feature representation. This fusion step allows the network to leverage information from both modalities. This fusion step creates a unified feature space where information from both modalities coexists. It allows the neural network to learn complex relationships between text and image data, enabling it to make more informed predictions about

human behavior.

4. Convolutional Layers: Convolutional layers are responsible for extracting meaningful features from the input data. They apply a set of learnable filters to the input image, convolving them across the image's spatial dimensions. Each filter learns to recognize different patterns, such as edges or textures, in the image. Adding more convolutional layers allows the network to capture increasingly complex features.

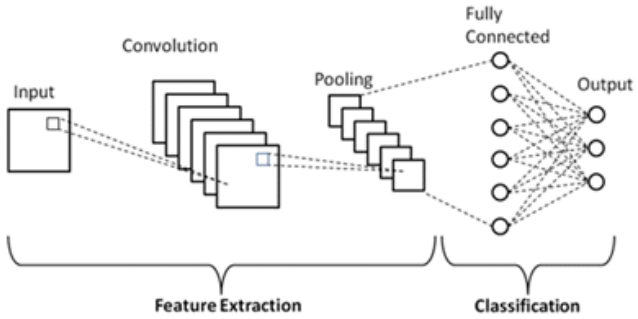

Figure 5: Architecture of a Convolutional Neural Network (CNN) with convolutional layers, pooling layers, and fully-connected (FC) layers

5. Max-Pooling Layers: Max-pooling layers are often used after convolutional layers. They reduce the spatial dimensions of the feature maps produced by convolution. Max-pooling operates by selecting the maximum value from a small region of the feature map, effectively down-sampling the data. This reduces computational complexity and focuses on the most relevant features.

6. Flatten Layer: After convolution and pooling, the feature maps are flattened into a one-dimensional vector. This transformation is essential to connect the convolutional layers to the fully connected layers. It preserves the learned features while preparing them for further processing.

7. Fully Connected Layers: Fully connected layers are traditional neural network layers where every neuron is connected to every neuron in the previous and subsequent layers and comprises the weights and biases together with the neurons. These layers make predictions based on the extracted features.

8. Dropout: Dropout layers are used during training to mitigate overfitting (when a model performs well on training data but not on new data), a common issue in deep learning. They randomly "drop out" a fraction of neurons during each training iteration, preventing the network from relying too heavily on any specific neuron. This encourages the network to learn more robust and generalizable features.

9. Output Layer: The output layer is the final layer of the network and typically contains one neuron per class in a classification task. The activation function in this layer depends on the task, but for multi-class classification, softmax is commonly

used. It computes the class probabilities for the input data.

10. Compile the Model: Compilation involves specifying the optimizer, loss function, and evaluation metrics. The optimizer updates the network's weights during training. The loss function quantifies how well the model's predictions match the actual data. Evaluation metrics, such as accuracy, provide insights into the model's performance during training and testing.

These components collectively define the architecture and functionality of a CNN, making it capable of learning and extracting valuable features from input data for various tasks, including image classification and behavioral data extraction.

4.2 Training the convolutional neural network

The training process of a convolutional neural network (CNN) begins with initializing random weights. During training, the CNN is provided with a large dataset of human behavioral data, where each data point is labeled with its corresponding behavioral category (e.g., walking, running, sitting). The CNN processes each behavioral data point, initially assigning random weights, and compares the output with the actual behavioral category label. If the predicted output does not match the labeled category, the CNN makes adjustments to its weights through a technique known as backpropagation. This iterative process optimizes the network's performance, gradually improving its accuracy in classifying behavioral patterns.

Each iteration through the entire dataset, termed an *epoch*, allows the CNN to refine its weights further, incrementally enhancing its ability to correctly classify behavioral data. As training progresses, the adjustments to the weights become smaller, indicating improved accuracy. Following the training phase, the CNN's performance is evaluated using a separate test dataset consisting of labeled behavioral data points not used during training. This evaluation assesses the CNN's ability to generalize its learned patterns to new, unseen behavioral data. If the CNN demonstrates high accuracy on the training dataset but performs poorly on the test dataset, it may indicate overfitting. Overfitting occurs when the model memorizes patterns specific to the training data rather than learning generalizable features, often due to a limited dataset size. So, the training process of the CNN involves iterative adjustments to its weights based on comparisons between predicted and actual behavioral categories, ultimately resulting in a model capable of accurately classifying human behavioral data. Evaluation using a separate test dataset ensures the model's ability to generalize its learned patterns to unseen data, mitigating the risk of overfitting.

In this research, an ensemble learning-based approach is introduced to enhance the recognition of human behaviors. Employing Weka 3, a powerful tool, facilitated the seamless combination of multiple models, yielding significant improvements in performance. Ensemble learning, at its core, harnesses the collective power of multiple learning algorithms to achieve superior results. Ensemble learning operates on the principle that the collaboration of several weak classifiers can yield a robust and

accurate strong classifier. It mitigates the impact of individual classifiers' errors by allowing them to collectively refine predictions. The ensemble learning approach can be categorized into two main types: homogeneous and heterogeneous ensemble learning. In the homogeneous ensemble, all individual learners share the same base model, while in the heterogeneous ensemble, diverse individual learners contribute their unique expertise.

In the experiments, a diverse ensemble was assembled comprising four distinct learners, each with its unique strengths. These learners encompassed:

- AdaBoost: AdaBoost, a widely recognized boosting algorithm, focuses on refining the accuracy of predictions by emphasizing previously misclassified data points.
- Random Forest: Leveraging the power of decision trees, Random Forest excels at capturing complex relationships within the data and offers robustness against overfitting.
- Bagging: Bagging, which stands for Bootstrap Aggregating, reduces variance and enhances stability by training multiple models on different subsets of the dataset and combining their predictions.
- Naïve Bayes: Naïve Bayes, a probabilistic classifier, provides valuable insights into behavioral patterns by modeling the likelihood of behaviors based on observed data.

The general structure of ensemble learning is to produce a group of individual learners and combine them together. Figure 6 shows the basic structure of ensemble learning.

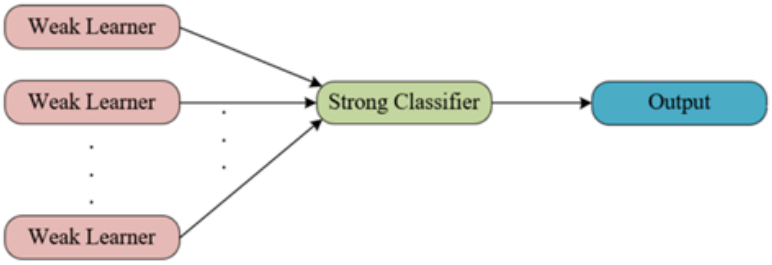

Figure 6: The basic structure of ensemble learning

Boosting method assists us to get a strong learner by combining a series of weak learners and integrating their learning ability. The adaptive boost (AdaBoost) adjusts the weight of samples base on the basis of previous learners, increases the proportion of samples that have been incorrectly classified, reduce the proportion of samples that have been correctly classified. The learners will focus on those samples that have been incorrectly classified. Finally, these learners are combined into a strong learner by weighting. Specifically, learners with high classification accu-

racy have higher weights, while learners with low classification accuracy have lower weights. Decision tree as the component of random forest, decision tree is a rapid and effective method with tree structure, in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. Random Forest is a classification algorithm that leverages the power of multiple decision trees for training and predicting samples. The algorithm's output is determined by considering the mode (most frequent) output category among the individual trees. This ensemble approach offers notable advantages, including high accuracy, scalability, and robustness against overfitting due to its use of bootstrap aggregation (bagging) and random feature selection during training. Random Forest is highly versatile and capable of handling both discrete and continuous data types, making it effective in diverse domains such as behavior recognition, medical diagnostics, and anomaly detection [19]. The Bootstrap Aggregating algorithm, often referred to as Bagging, constructs multiple weak learners independently. These individual learners operate in parallel, allowing for simultaneous training. Subsequently, their collective insights are harnessed to enhance the overall predictive performance. Bagging can be effectively combined with various classification and regression algorithms, resulting in improved accuracy and model stability. Importantly, this ensemble technique reduces result variance, mitigating the risk of overfitting. The Naïve Bayes method is rooted in Bayesian algorithms, operating under the assumption that, given the target value, the attributes are statistically independent of one another. In simpler terms, no single attribute variable exerts a disproportionately significant influence on decision-making, nor does any attribute variable exert an overly diminished influence. Despite this simplification, Naïve Bayes classifiers have shown remarkable effectiveness in real-world scenarios, especially where rapid classification is essential or datasets are high-dimensional [1]. This statistical approach facilitates efficient and effective decision-making processes, making it particularly valuable in various classification tasks such as text categorization, emotion recognition, and even early-stage behavior prediction in intelligent systems. Recent studies have extended the use of Naïve Bayes in ensemble systems, where it acts as a lightweight yet informative learner when combined with more complex models like Random Forest or Gradient Boosting Machines [8]. Its probabilistic nature complements deterministic models, enhancing the diversity of the ensemble and often leading to better generalization performance. Our ensemble learning approach, encompassing these diverse learners, significantly contributes to the effectiveness and reliability of our behavioral data extraction process, aligning with the core objectives of our research.

In the realm of human behavior recognition, traditional machine learning methods have predominantly relied on feature extraction techniques, with a particular emphasis on spatial information [32]. However, it's worth noting that spatial information can be influenced by external environmental factors. In contrast, a notable advancement in this field has been the introduction of Improved Dense Trajectories (IDT), which has made substantial contributions to human behavior recognition.

Deep learning approaches, on the other hand, have highlighted the significance of both spatial and temporal information in capturing intricate motion features. Consequently, our research incorporates Long Short-Term Memory (LSTM) networks to extract crucial temporal information from each video frame [10]. The inclusion of LSTM allows us to capture the dynamic aspects of behavior recognition, complementing the spatial information. Figure 7 illustrates the fundamental architecture of LSTM within our study. This integration of LSTM into our methodology underscores our commitment to capturing not only spatial but also temporal nuances, ultimately enhancing the accuracy and effectiveness of human behavior recognition.

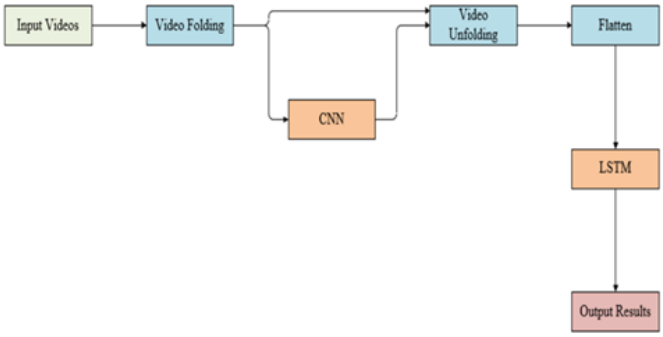

Figure 7: The basic LSTM architecture for human behavior recognition

By combining Convolutional Neural Networks (CNN) with LSTM, our model achieves exceptional accuracy in human behavior recognition. CNN processes each video frame independently, while LSTM effectively captures temporal dependencies. This synergy enhances our model's capability to recognize complex behavioral patterns. CNN acts as a feature extractor, and its output seamlessly feeds into LSTM for comprehensive analysis, resulting in high accuracy.

5 Experimental Setup and Evaluation

5.1 Hyperparameter Configuration

The performance of the proposed framework depends on carefully selected hyperparameters to ensure convergence, stability, and generalization. The CNN-LSTM architecture was trained using empirically optimized settings, as summarized in Table 2.

The Adam optimizer was selected due to its adaptive learning capability and faster convergence. Dropout regularization was incorporated to reduce overfitting, particularly given the variability in human behavioral data. The number of epochs

Parameter	Value
Learning Rate	0.001
Batch Size	32
Number of Epochs	50
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Dropout Rate	0.5
Activation Function	ReLU (hidden), Softmax (output)
LSTM Units	128
Number of CNN Layers	4

Table 2: Hyperparameter Settings.

and batch size were chosen based on validation performance to balance training time and model accuracy.

5.2 Evaluation Metrics

To assess the effectiveness of the proposed model, standard classification metrics were employed. These metrics provide a comprehensive evaluation of the model's ability to correctly recognize human behavioral data. The following metrics were used:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + PFN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Accuracy measures the overall correctness of the model, while precision and recall evaluate its ability to minimize false positives and false negatives, respectively. The F1-score provides a balanced measure, particularly useful when class distributions are uneven.

5.3 Baseline Models for Comparison

To validate the effectiveness of the proposed framework, comparisons were conducted with several baseline models, including both traditional and deep learning approaches:

- CNN-only model - captures spatial features
- LSTM-only model - captures temporal dependencies
- Support Vector Machine (SVM) - traditional classifier
- Random Forest (RF) - ensemble-based classical model

These baselines provide a comprehensive benchmark to evaluate improvements achieved by the proposed CNN-LSTM and ensemble framework.

5.4 Quantitative Results

The performance comparison across different models is presented in Table 3.

Model	Accuracy	Precision	Recall	F1-score
SVM	80.3	79.8	78.5	79.1
Random Forest	82.7	82.1	81.6	81.8
CNN	85.2	84.5	83.9	84.2
LSTM	87.1	86.4	85.8	86.1
CNN-LSTM (proposed)	91.6	91.2	90.8	91.0
Ensemble Model	**93.4**	**93.0**	**92.5**	**92.7**

Table 3: Performance Comparison of Models.

5.5 Performance Analysis

The results indicate that the proposed CNN-LSTM model significantly outperforms individual models by effectively capturing both spatial and temporal characteristics of human behavioral data. The CNN component extracts discriminative visual features, while the LSTM captures temporal dependencies across frames.

Furthermore, the ensemble model achieves the highest performance across all metrics. This improvement can be attributed to the complementary strengths of multiple classifiers, which enhance robustness and reduce model variance. Compared to traditional methods such as SVM and Random Forest, deep learning models demonstrate superior capability in handling complex, high-dimensional behavioral data.Overall, the results validate the effectiveness of the proposed framework in achieving accurate and reliable human behavior recognition.

6 Conclusion and Discussion

This research introduces a novel framework for representing, analyzing, and generating human behavioral data through the integration of deep learning techniques,

particularly Convolutional Neural Networks and Long Short-Term Memory (LSTM) models. The proposed methodology places strong emphasis on multimodal signal processing, enabling the effective modeling and synthesis of complex human behaviors such as postures, gestures, facial expressions, physiological signals, and emotional states. The ability to analyze behavior across multiple modalities in real-time stands out as a core strength of this approach.

One of the pivotal contributions of this study lies in its ensemble learning architecture, which leverages the complementary strengths of classifiers such as AdaBoost, Random Forest, Bagging, and Naïve Bayes. This ensemble strategy not only improves classification accuracy but also enhances the generalizability of the model across varied contexts. Furthermore, the development and use of custom high-resolution datasets address practical challenges seen in many publicly available datasets, allowing the models to perform more reliably under realistic scenarios.

The impact of this research extends across various domains. In healthcare, the framework can support continuous patient monitoring and emotion detection, while in education, it can be used to assess engagement and deliver personalized content. Entertainment industries can leverage the synthesis capabilities for emotionally resonant content generation in virtual environments and gaming. These applications underscore the potential of this work to redefine human-computer interaction, making it more empathetic, responsive, and context-aware.

However, the study is not without its challenges. The integration of diverse data modalities inherently increases system complexity, requiring careful tuning and extensive preprocessing. Manual labeling of regions of interest (ROIs), while effective, is labor-intensive and may introduce human error or bias. In addition, the computational demands of deep learning architectures present scalability challenges, particularly when considering real-time deployment on edge devices or in low-resource environments.

Despite these limitations, the research paves the way for future exploration in behavior modeling. Future work could include the adoption of attention mechanisms and graph neural networks for improved pattern recognition, the automation of data labeling through weak supervision or active learning, and the adaptation of models for edge computing environments. Longitudinal analysis of behavioral data and integration with physiological and contextual sensors could also provide deeper insights into dynamic human behavior over time.

By constructing a holistic framework that not only extracts but also synthesizes multimodal behavioral signals, this study contributes a significant step forward in decoding the richness of human expression. Our model architecture, grounded in CNN and LSTM layers, achieves high accuracy and can be readily adapted for use across disciplines such as psychology, robotics, and smart systems. It opens new pathways for researchers and practitioners to design intelligent systems capable of understanding and responding to human behavior in a more authentic, human-centric manner.

In conclusion, this work represents a critical fusion of deep learning and behavioral science, laying the foundation for a future where artificial systems can interpret, generate, and interact with human behaviors in ways that are intuitive, ethical, and effective. It underscores the transformative power of deep learning in shaping a digital ecosystem that is empathetic, adaptive, and fundamentally aligned with human needs.

7 Future Work and Limitations

In considering future directions for this research, several promising avenues emerge that could significantly advance its impact and applicability. First and foremost, the integration of advanced deep learning techniques, such as attention mechanisms or graph neural networks, holds great potential for enhancing the capability to capture intricate patterns in human behavior data. These techniques offer improved temporal modeling and attention to relevant features, potentially leading to heightened accuracy and robustness in behavior recognition tasks. Additionally, exploring strategies for real-time processing and deployment of the developed models in practical applications, such as healthcare, surveillance, and human-computer interaction, could vastly broaden their utility and real-world impact. This involves not only optimizing model architecture and algorithms for efficiency and scalability but also carefully considering hardware constraints for deployment on edge devices. Furthermore, investigating methods for effectively fusing multimodal data sources, including text, image, and sensor data, could provide a more comprehensive understanding of human behavior, thus improving recognition accuracy. Additionally, adapting the proposed methodology to specific domains like healthcare or security through customization of models and algorithms could yield more targeted and effective solutions. Finally, conducting longitudinal studies to analyze changes and trends in human behavior over time could offer valuable insights into behavioral patterns, deviations, and interventions, further enriching the research's contributions to the field. Addressing these future work areas holds great promise for advancing the research's impact and fostering the development of more accurate and reliable systems for behavior recognition across diverse applications.

Data Availability Statement: Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Competing Interests: This study has no conflicts of interest and is not funded by any Organization/Institution. All authors have participated in conception and design, anal-lysis and interpretation of the data, drafting the article or revising it critically for important intellectual content, and approval of the final version.

Ethical and Informed Consent for Data Used: Not Applicable.

References

[1] M. Adhikari & A. Munusamy, *ICovidCare: Intelligent health monitoring framework for COVID-19 using ensemble random forest in edge networks*, Internet of Things, **14**, (2021), 100385. DOI: https://doi.org/10.1016/j.iot.2021.100385.

[2] M. A. Akbar, A. A. Khan, S. Mahmood, S. Rafi & S. Demi, *Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges*, Software: Practice and Experience, **54**(9), (2024), 1621–1650. DOI: https://doi.org/10.1002/spe.3216.

[3] A. Ali, W. Samara, D. Alhaddad, A.Ware & O.A. Saraereh, *Human activity and motion pattern recognition within indoor environment using convolutional neural networks clustering and naive bayes classification algorithms*, Sensors, **22**(3):1016, (2022). DOI: https://doi.org/10.3390/s22031016.

[4] N. Alruwais & M. Zakariah, *Student-engagement detection in classroom using machine learning algorithm*, Electronics, **12**(3), (2023), 731. DOI: https://doi.org/10.3390/electronics12030731.

[5] T. Baltrušaitis, C. Ahuja & L.P. Morency, *Multimodal machine learning: A survey and taxonomy*, IEEE Trans. Pattern Anal. Mach. Intell., **41**(2), (2019), 423–443. DOI: https://doi.org/10.1109/TPAMI.2018.2798607.

[6] P. Bhardwaj, P.K. Gupta, H. Panwar, M.K. Siddiqui, R. Morales-Menendez & A. Bhaik, *Application of deep learning on student engagement in e-learning environments*, Computers & Electrical Engineering, **93**, (2021), 107277. DOI: https://doi.org/10.1016/j.compeleceng.2021.107277.

[7] F. Caba Heilbron, V. Escorcia, B. Ghanem & J. Carlos Niebles, *ActivityNet: A large-scale video benchmark for human activity understanding*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, (2015), 961-970. DOI: https://doi.org/10.1109/CVPR.2015.7298698.

[8] C.W. Chang, C.Y. Chang & Y.Y. Lin, *A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection*, Multimedia Tools and Applications, **81**(9), (2022), 11825–11843. DOI: https://doi.org/10.1007/s11042-021-11887-9.

[9] D.C. Ciresan, U. Meier, L.M. Gambardella & J. Schmidhuber, *Convolutional neural network committees for handwritten character classification*, in Proc. Int. Conf. Document Anal. Recognit., Sep., (2011), 1135-1139. DOI: https://doi.org/10.1109/ICDAR.2011.229.

[10] P. Datta & R. Rohilla, *An autonomous and intelligent hybrid CNN-RNN-LSTM based approach for the detection and classification of abnormalities in brain*, Multimedia Tools and Applications, **83**(21), (2024), 60627-60653. DOI: https://doi.org/10.1007/s11042-023-17877-3.

[11] A. Graves, *Supervised Sequence Labelling*, Springer, 2012, 5-13. DOI: https://api.semanticscholar.org/CorpusID:60085539.

[12] A. Graves, A.R. Mohamed & G. Hinton, *Speech recognition with deep recurrent neural networks*, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May, (2013), 6645-6649. DOI: https://doi.org/10.1109/ICASSP.2013.6638947.

[13] K. Gupta, A. Singh, S.R. Yeduri, M.B. Srinivas & L.R. Cenkeramaddi, *Hand gestures recognition using edge computing system based on vision transformer and lightweight CNN*, J. Ambient Intell. Humanized Comput., **14**(3), (2023), 2601–2615. DOI: https://doi.org/10.1007/s12652-022-04506-4.

[14] Z. Hua, Z. Wang, X. Xu, X. Kong & H. Song, *An effective PoseC3D model for typical action recognition of dairy cows based on skeleton features*, Computers and Electronics in Agriculture, **212**, 108152, (2023). DOI: https://doi.org/10.1016/j.compag.2023.108152.

[15] S. Indolia, A.K. Goswami, S.P. Mishra & P. Asopa, *Conceptual understanding of convolutional neural network—a deep learning approach*, Procedia Computer Science, **132**, (2018), 679–688. DOI: https://doi.org/10.1016/j.procs.2018.05.069.

[16] M.M. Islam, S. Hassan, S. Akter, F.A. Jibon & M. Sahidullah, *A comprehensive review of predictive analytics models for mental illness using machine learning algorithms*, Healthcare Analytics, **6**, (2024), 100350. DOI: https://doi.org/10.1016/j.health.2024.100350.

[17] N. Jaouedi, N. Boujnah & M.S. Bouhlel, *A new hybrid deep learning model for human action recognition*, J. King Saud Univ.–Comput. Inf. Sci., **32**(4), (2020), 447–453. DOI: https://doi.org/10.1016/j.jksuci.2019.09.004.

[18] S. Ji, W. Xu, M. Yang & K. Yu, *3D convolutional neural networks for human action recognition*, IEEE Trans. Pattern Anal. Mach. Intell., **35**(1), (2012), 221–231. http://researchonline.ljmu.ac.uk/id/eprint/9438/.

[19] A. Khan, A. Sohail, U. Zahoora & A.S. Qureshi, *A survey of the recent architectures of deep convolutional neural networks*, Artif. Intell. Rev., **53**(8), (2020), 5455–5516. DOI: https://doi.org/10.1007/s10462-020-09825-6.

[20] J. Li, K. Jin, D. Zhou, N. Kubota & Z. Ju, *Attention mechanism-based CNN for facial expression recognition*, Neurocomputing, **411**, (2020), 340–350. DOI: https://doi.org/10.1016/j.neucom.2020.06.014.

[21] J. Liu et al., *NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding*, IEEE Trans. Pattern Anal. Mach. Intell., **42**(10), (2020) 2684–2701. DOI: https://doi.org/10.1109/TPAMI.2019.2916873.

[22] Y. Liu, H. Zhang, Y. Zhan, Z. Chen, G. Yin, L. Wei & Z. Chen, *Noise-resistant multimodal transformer for emotion recognition*, International Journal of Computer Vision, (2024), 1-21. DOI: https://doi.org/10.1007/s11263-024-02304-3.

[23] F. Luo, S. Khan, B. Jiang & K. Wu, *Vision transformers for human activity recognition using WiFi channel state information*, IEEE Internet of Things Journal, (2024). DOI: https://doi.org/10.1109/JIOT.2024.3375337.

[24] S. Marcos-Pablos & F.J. García-Peñalvo, *Emotional intelligence in robotics: A scoping review*. In New Trends in Disruptive Technologies,Tech Ethics and Artificial Intelligence: The DITTET Collection 1. Springer International Publishing.(2022)(pp. 66–75) DOI: https://doi.org/10.1007/978-3-030-87687-6_7.

[25] S. Mekruksavanich & A. Jitpattanakul, *Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition*, Sci. Rep., **13**(1), (2023), 12067. DOI: https://doi.org/10.1038/s41598-023-39080-y.

[26] A. van den Oord et al., *WaveNet: A generative model for raw audio*, arXiv preprint arXiv:1609.03499, (2016). DOI: https://doi.org/10.48550/arXiv.1609.03499.

[27] S. Peng, S. Sun & Y.D. Yao, *A survey of modulation classifi- cation using deep learning: Signal representation and data preprocessing*, IEEE Transactions on Neural Networks and Learning Systems, **33**(12), (2021), 7020–7038. DOI: https://doi.org/10.1109/TNNLS.2021.3085433.

[28] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar & K. Muhammad, *Human behavior understanding in big multimedia data using CNN based facial expression recognition*, Mobile Netw. Appl., **25**, (2020), 1611–1621. DOI: https://doi.org/10.1007/s11036-019-01366-9.

[29] W. Sheng, P. Shan, S. Chen, Y. Liu & F. E. Alsaadi, *A niching evolutionary algorithm with adaptive negative correlation learning for neural network ensemble*, Neurocomputing, **247**, (2017), 173-182. DOI: https://doi.org/10.1016/j.neucom.2017.03.055.

[30] X. Shao, R. Niu, X. Shao, J. Gao, Y. Shi, Z. Jiang & Y. Wang, *Application of dual-stream 3D convolutional neural network based on 18 F-FDG PET/CT in distinguishing benign and invasive adenocarcinoma in ground-glass lung nodules*, EJNMMI Physics, **8**, (2021),1–13. DOI: https://doi.org/10.1186/s40658-021-00423-1.

[31] P. Tarnowski, M. Kołodziej, A. Majkowski & R.J. Rak, *Emotion recognition using facial expressions*, Procedia Computer Science, **108**, (2017), 1175–1184. DOI: https://doi.org/10.1016/j.procs.2017.05.025.

[32] T. Vaikunta Pai et al., *DKCNN: Improving deep kernel convolutional neural network-based COVID-19 identification from CT images of the chest*, Journal of X-Ray Sci. Technol., Preprint, **32**(1), (2024), 1–18. DOI: https://doi.org/10.3233/XST-230424.

[33] A. Vaswani et al., *Attention is all you need*, Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), (2017). DOI: https://doi.org/10.48550/arXiv.1706.03762.

[34] J. Wang, Y. Chen, S. Hao, X. Peng & L. Hu, *Deep learning for sensor- based activity recognition: A survey*, Pattern Recognit. Lett., **119** (2019), 3-11. DOI: https://doi.org/10.1016/j.patrec.2018.02.010.

[35] Q. Xu et al., *Scene image and human skeleton-based dual-stream human action recognition*, Pattern Recognit. Lett., **148**, (2021),136–145. DOI: https://doi.org/10.1016/j.patrec.2021.06.003.

[36] H. Yang et al., *Asymmetric 3D convolutional neural networks for action recognition*, Pattern Recognit., **85**, (2019), 1–12. DOI: https://doi.org/10.1016/j.patcog.2018.07.028.

[37] Z. Yu & W.Q. Yan, *Human action recognition using deep learning methods*, in Proc. 35th Int. Conf. Image and Vision Comput. New Zealand (IVCNZ), Nov., (2020), 1-6. DOI:https://doi.org/10.1109/IVCNZ51579.2020.9290594.

[38] E.M. Younis, S.M. Zaki, E. Kanjo & E.H. Houssein, *Evaluating ensemble learning methods for multimodal emotion recognition using sensor data fusion*, Sensors, **22**(15), (2022), 5611. DOI: https://doi.org/10.3390/s22155611.

[39] Q. Zhang & W.Q. Yan, *Currency detection and recognition based on deep learning*, Proc. 15th IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), Nov., (2018), 1-6. DOI:https://doi.org/10.1109/AVSS.2018.8639124.

Vaikunta Pai T, ORCID: 0000-0001-6100-9023
Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT),
Department of Information Science and Engineering, Nitte, India.
e-mail: vaikunthpai@gmail.com

Manjula Mallya M, ORCID: 0009-0005-8812-6912
Government First Grade College for Women, Mangalore, India.
e-mail: manjulamallya88@gmail.com

Ramona Birau, ORCID: 0000-0003-1638-4291
Eugeniu Carada Doctoral School of Economic Sciences,
University of Craiova, Craiova, România.
e-mail: ramona.f.birau@gmail.com

Nethravathi P S, ORCID: 0000-0001-5447-8673
Department of Computer Science and Engineering, Shree Devi Institute of Technology,
Mangalore, India.
e-mail: nethrakumar590@gmail.com

Virgil Popescu, ORCID: 0009-0002-0269-2541
University of Craiova, Faculty of Economics and Business Administration, Craiova, România.
e-mail: virgil.popescu@vilaro.ro

Iuliana Carmen Bărbăcioru - Corresponding author, 0000-0001-7329-9590
Faculty of Engineering, Constantin Brâcuşi University of Târgu Jiu, Gorj, România.
e-mail: cbarbacioru@gmail.com

Pramod Vishnu Naik, ORCID: 0009-0002-6551-964x
Software Development Engineer, MResult Services Private Limited, Mangalore, India.
e-mail: pammunaik92@gmail.com

License

Received: January 07, 2026; Accepted: April 08, 2026; Published: April 09, 2026.