A New Hybrid Algorithm for Vision-Based Sleep Posture Analysis Integrating CNN, LSTM and MediaPipe

Apichaya Nimkoompai, Puwadol Sirikongtham*
Faculty of Information Technology, Thai-Nichi Institute of Technology, Thailand

***Abstract*—Sleep posture is a critical factor affecting sleep quality and long-term health, particularly for the elderly and patients with chronic conditions. This research proposes a novel hybrid algorithm for real-time, vision-based sleep posture analysis by integrating Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and MediaPipe pose estimation. The primary objective is to accurately classify the four main sleep postures—supine, left lateral, right lateral, and prone—while incorporating an automated alert system for risky behaviors, such as maintaining a prone position for over 15 minutes or remaining in any static posture for more than 2 hours. The system processes video input through a streamlined pipeline: MediaPipe first extracts 3D body keypoints, which are then fed into a CNN for spatial feature extraction, followed by an LSTM network to model temporal dependencies across frames. Evaluated on a dataset of 280 video samples from 20 participants under both daytime and nighttime conditions, the model achieved an accuracy of 96.4% in daylight and 92.8% in low-light environments, demonstrating robust performance across varying illumination. Comparative analysis confirmed its superiority over existing methods, such as depth-based CNN or pressure-sensor models. The study concludes that the proposed hybrid system offers a practical, non-invasive, and highly accurate solution for continuous sleep monitoring, with significant potential for deployment in smart healthcare and remote elderly care applications.**

***Keywords*—Sleep posture detection; MediaPipe; CNN; LSTM; real-time monitoring; pose estimation**

I. Introduction

Sleep posture analysis using vision-based systems is regarded as a significant technology in the healthcare sector, particularly for the elderly and patients requiring special monitoring. Although improper sleep postures are not direct causes of severe diseases, they can act as contributing factors that weaken the body, potentially leading to musculoskeletal disorders, spinal problems, respiratory issues, and gastroesophageal reflux disease if accumulated over a long period. Therefore, maintaining correct sleep posture is not only essential for preventing illness but also for prolonging the body's functionality, enhancing quality of life, and reducing the risk of future diseases, making it a matter of great importance.

In this research, a "Novel hybrid algorithm for vision-based sleep posture analysis by integrating CNN, LSTM, and MediaPipe" was developed, providing a solution for real-time sleep posture monitoring.

Despite these advancements, a significant research gap remains in developing a vision-based system that simultaneously achieves high accuracy, robust performance in varying lighting conditions, real-time processing on affordable hardware, and built-in privacy protection. Many existing systems rely on specialized sensors like depth cameras or pressure mats, which increase cost and complexity [14, 15]. Other approaches explore the use of thermal imaging for monitoring patients with pressure ulcer risk, focusing on long-term monitoring and image-based posture detection [4, 5, 6]. Other vision-based approaches that process raw video frames raise substantial privacy concerns and require significant computational resources, making them less suitable for continuous, in-home monitoring [3]. Furthermore, many solutions lack an integrated temporal model capable of not only classifying postures accurately but also continuously monitoring posture duration to proactively alert for health risks. This study aims to bridge this gap by proposing a novel hybrid algorithm that effectively addresses these limitations through a strategic integration of MediaPipe, CNN, and LSTM technologies.

The proposed system possesses advanced capabilities in detecting and classifying primary sleep postures, as well as identifying risky postures that may lead to health problems, for example, prolonged prone positions that can cause breathing difficulties, or maintaining the same posture for more than two hours, which may result in pressure ulcers. These functions are enabled through the integration of three key technologies: MediaPipe for 3D pose estimation, CNN for spatial feature extraction, and LSTM for temporal sequence analysis of movements.

A distinctive advantage of this system is its ability to function effectively even under low-light conditions, which are typical during sleep, while maintaining real-time processing performance. This research was tested and validated using data collected from 20 participants aged between 19 and 21 years, with heights ranging from 1.55 to 1.80 meters and weights between 45 and 90 kilograms. The dataset was designed to include variations in body mass index (BMI) and sleep attire to enhance model generalizability. Video recordings were captured during both daytime and nighttime under controlled and uncontrolled lighting scenarios. The experimental results demonstrated superior performance compared to conventional methods, with an average accuracy of 94.6% and a significantly reduced false alert rate.

*Corresponding author.

This research introduces a novel hybrid algorithm for vision-based sleep posture analysis through a purpose-built integration of CNN, LSTM, and MediaPipe. The key innovation lies in the specialized fusion strategy that processes 3D skeletal keypoints from MediaPipe through a 1D CNN for spatial feature extraction, followed by an LSTM network that performs dual functions of posture sequence classification and direct temporal monitoring for risk alerts. This integrated approach is specifically optimized for real-world sleep monitoring constraints, including low-light robustness and privacy preservation through on-device processing of anonymized skeletal data only.

This innovation not only addresses technical challenges in image processing and deep learning but also has high potential for practical application in healthcare facilities or elderly residences to help prevent and monitor health issues caused by incorrect sleep postures, without the need for invasive contact sensors that may disturb users.

II. Theories and Related Research

A. Sleep Postures Affecting Health and Risk Behaviors

Sleeping in inappropriate postures for extended periods may negatively affect health, particularly among the elderly and patients with specific health conditions [1]. The primary sleep postures considered to pose health risks can be categorized into two main types: 1) prone posture, which may compress internal organs and the respiratory system, and 2) maintaining the same posture continuously for more than two hours, which may cause pressure ulcers and could indicate sleep-related problems.

The mechanisms through which improper sleep postures affect health involve several factors, including tissue and circulatory compression, restricted lung expansion, and stress exerted on joints and muscles [2]. Medical research has found that prone sleeping is associated with a higher risk of sleep apnea, while remaining in the same posture for an extended duration increases the likelihood of developing pressure ulcers, especially among the elderly with reduced blood circulation.

The sleep posture analysis system in this study was designed to detect and monitor three primary risk behaviors: maintaining a prone posture for more than 15 minutes and remaining in the same posture continuously for over two hours. These criteria were established based on medical evidence, which indicates that prolonged prone sleeping is linked to an increased risk of sleep apnea, while immobility for extended periods is a significant risk factor for pressure ulcers, particularly in patients or elderly individuals with poor blood circulation. The system is therefore capable of issuing real-time alerts, enabling caregivers to provide timely assistance. Moreover, research conducted by Kim investigated the effects of different sleep postures on intraocular pressure (IOP) and ocular perfusion pressure (OPP). Participants were instructed to sleep in supine, left lateral, right lateral, and prone positions [16]. The findings revealed that the prone position resulted in the highest IOP, whereas the supine position produced the highest OPP. This proved that sleep posture significantly influences intraocular pressure and ocular blood flow, with the prone posture potentially being a risk factor for eye diseases. These studies indicated that sleep posture has substantial impacts on various aspects of health.

B. MediaPipe

MediaPipe is a real-time human pose estimation technology developed by Google Research, specifically designed for mobile devices and edge computing systems. This technology consists of two main components: a body region detector and a 3D keypoint estimator that identifies 33 keypoints on the human body [7]. A key advantage of MediaPipe lies in its high benchmarked processing speed, which can exceed 30 frames per second on typical smartphones, combined with high accuracy even for complex poses, and efficient 3D pose estimation capabilities [8]. The use of 3D data (X, Y, Z coordinates) enables the system to distinguish complex postures that may appear similar in 2D images, such as differentiating between a slightly twisted supine posture and a lateral posture, thereby significantly improving classification accuracy.

In this research, MediaPipe was utilized as the fundamental component for extracting body keypoints from video frames, converting pose data into 3D formats, and forwarding the processed data to the CNN-LSTM system for further analysis [8]. Compared with other technologies like OpenPose and PoseNet, MediaPipe proves clear advantages in speed, lower system resource usage, and accuracy in real-world environments [7]. Although it has some limitations in detection when parts of the body are occluded or under very low lighting conditions, this study developed image quality enhancement and data imputation techniques to address these issues.

The data obtained from MediaPipe undergoes further processing before practical use, including coordinate calibration, calculation of angles between keypoints, and sequence creation for input into the LSTM network. Overall, MediaPipe is the core technology enabling our system to effectively detect and track sleep postures in real-time, delivering fast and accurate performance [8], making it highly suitable for application in sensitive and reliable elderly health monitoring systems.

C. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) are neural network architectures specifically designed for processing grid-structured data, such as digital images [9]. They excel at automatically extracting spatial features through three main mathematical operations: convolution, pooling, and fully connected layers.

The convolution process employs learnable filters or kernels to detect various levels of image features, from edges and basic shapes to increasingly complex patterns at deeper layers [9]. Each layer produces feature maps that represent the response to specific spatial characteristics of the image. Pooling operations reduce data dimensionality and improve robustness to small spatial variations of objects. Fully connected layers translate the extracted features into classification results.

A limitation of CNN operating alone is that it analyzes only single images, which can lead to difficulties in distinguishing similar postures, such as supine and lateral postures seen from similar camera angles [10]. CNN's strengths include weight sharing in convolutional layers, which reduces the number of parameters compared to traditional networks, and hierarchical feature learning that builds from simple to complex features, making CNN highly suitable for image analysis tasks.

In this research, CNN was used in combination with MediaPipe to extract deep spatial features from body landmark points before passing data to LSTM for temporal relationship analysis [10]. This integration enhances the overall system's performance in real-time sleep posture classification.

D. Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory (LSTM) Networks are a type of Recurrent Neural Network (RNN) designed to address the challenge of learning temporal dependencies in long sequential data. Traditional RNNs often suffer from gradient vanishing or gradient explosion when processing long sequences [11]. LSTM overcomes these limitations through a memory cell mechanism and controlling gates that allow the network to selectively remember or forget information as needed.

The main structure of LSTM consists of three gates: 1) the Forget Gate, which determines what information from the previous state is retained or discarded; 2) the Input Gate, which controls the update of new information into the memory cell; and 3) the Output Gate, which decides what information is output to the next state [11]. Each gate uses activation functions such as Sigmoid and Tanh to regulate values within appropriate ranges. This mechanism enables LSTM to effectively learn long-term dependencies in time-series data.

A limitation of standalone LSTM is the lack of inherent spatial representation, as it relies on input vectors or embeddings [10]. In this research, LSTM was employed to analyze sequences of sleep postures derived from body landmark data obtained through MediaPipe and processed by CNN. It processes dynamic posture changes between consecutive video frames, improving accuracy in posture classification and detecting risky behaviors such as prolonged static postures or abnormal posture transitions. The integration of LSTM with CNN enables comprehensive spatial-temporal processing, resulting in a robust hybrid model for real-time sleep posture analysis.

E. Integration of MediaPipe, CNN, and LSTM

The real-time sleep posture analysis system in this research relies on integrating the strengths of three core technologies: MediaPipe, CNN, and LSTM to form a highly efficient hybrid algorithm, where each component contributes sequentially to the processing workflow [8], [10], [11].

Initially, MediaPipe acts as the pose estimation module, detecting and estimating human body keypoints from video frames using a lightweight CNN architecture. It provides 3D landmarks of the body with high speed and suitable accuracy for real-time processing [7], [8]. MediaPipe's advantage lies in its ability to operate on standard devices without reliance on high-performance servers, reducing raw image complexity into structural landmark points that represent sleep postures [7].

Next, CNN receives the landmark data from MediaPipe to extract meaningful spatial features specific to each sleep posture [9]. The CNN architecture is designed to learn deep patterns such as the spatial relationships among the head, shoulders, arms, and legs associated with supine, lateral, or prone postures. Convolutional layers detect key features like joint angles and body arrangement.

Finally, LSTM analyzes the temporal sequence of features derived from CNN to detect dynamic posture changes and evaluate risky behaviors, such as prolonged static postures or abnormal rotations [11]. LSTM plays a critical role in capturing temporal dependencies between consecutive frames, reducing errors from frame-by-frame classification and improving accuracy in detecting complex sleep states, such as transitions from lateral to curled postures.

Together, CNN + LSTM integrate "shape" and "sequential behavior" for more accurate posture recognition [10]. The combined workflow forms an end-to-end system processing from image input to alert generation.

MediaPipe simplifies raw data, CNN extracts spatial features to identify "which posture this is", and LSTM interprets the temporal context to understand "whether the postures in the video sequence are continuous or changing", such as stable sleep, posture shifts, or turning over. This integrated approach enables comprehensive and highly accurate sleep posture analysis at real-time speeds, meeting practical requirements for effective elderly healthcare monitoring.

F. Related Research on Sleep Posture Detection

In recent years, numerous studies have focused on developing sleep posture detection systems using deep learning technologies such as CNN and LSTM to improve classification accuracy and detect health-risk behaviors. These studies commonly utilize video data or sensors to analyze postures in real-time and address practical limitations such as low lighting or body occlusion by blankets.

One related study by Tang et al. developed an intelligent sleep posture detection system using CNN to analyze pressure data from a smart mat [12]. The system classified four main postures with high accuracy and included alert functions when risky postures, such as prolonged prone sleeping, were detected. It was suitable for home use and operated non-invasively without disturbing the user. Experimental results showed accuracy exceeding 90% in real-world environments. However, the system presented in this research offers the practical advantage of being completely non-contact, further minimizing disturbance to users' sleep.

Another study by Chung et al. [13] evaluated the performance of Gated Recurrent Neural Networks (GRU), a technology related to LSTM, for analyzing temporal sequence data. The work focused on classification and prediction of sequences, which can be applied to detect posture changes such as prolonged immobility or abnormal movements during sleep. Results indicated that the gate mechanism improved processing efficiency for dynamic data, making them suitable for elderly health monitoring systems, though training and deployment can be complex.

Current research still faces three main challenges: 1) managing body occlusions, 2) enhancing processing speed for real-time applications, and 3) increasing accuracy in classifying similar postures. This study proposes a novel approach integrating the strengths of MediaPipe, CNN, and

LSTM to address these limitations, especially in terms of speed and classification accuracy.

III. Research Methodology

A. Novelty in Architecture Integration

The integration of MediaPipe, CNN, and LSTM in this work represents a deliberate architectural design innovation targeting the specific challenges of sleep monitoring. Our main contribution lies in a specialized fusion strategy that differs fundamentally from existing approaches in three key aspects:

1) Structured data-centric pipeline: Unlike conventional vision-based models that process raw pixels, we introduce MediaPipe as a pre-processing abstraction layer. This converts high-dimensional video data into low-dimensional 3D skeletal keypoints before deep learning analysis, creating a privacy-preserving and computationally efficient foundation.

2) Dual-function temporal modeling: The LSTM network in our architecture is designed for a dual purpose: it not only enhances posture classification accuracy through temporal sequence analysis but also directly powers the risk alert mechanism by continuously tracking posture duration. This tight coupling of classification and risk assessment within a single temporal model is a novel approach for sleep monitoring systems.

3) End-to-end optimization for real-world deployment: The entire architecture is optimized for practical constraints, including robustness to low-light conditions via IR camera compatibility, and real-time performance on consumer-grade hardware through efficient skeletal data processing.

This targeted integration strategy results in a system that is not only accurate but also practical, efficient, and privacy-conscious, specifically engineered for the demands of in-home, long-term sleep monitoring.

B. System Architecture

The proposed system receives real-time video input, and the output consists of sleep posture predictions along with recommendations for posture adjustment in terms of both angle and position. The system comprises three main stages: 1) keypoint extraction, 2) sleep posture prediction, and 3) feedback for posture adjustment.

In the keypoint extraction stage, MediaPipe technology is utilized to detect and identify keypoints from the video frames. These points are then utilized for posture analysis in the subsequent stage. In the posture prediction stage, a hybrid architecture integrating CNN and LSTM is applied to classify sleep postures and assess whether the user is in a correct or improper posture, based on the spatial and temporal relationships among the detected keypoints.

In the final stage, the system provides real-time feedback for posture correction if abnormalities are detected, while also displaying the percentage of similarity to the correct posture. This feedback allows users to appropriately adjust their sleep posture, which is crucial for preventing potential health issues resulting from prolonged improper sleeping positions.

The system was designed to operate in an end-to-end manner, starting from video input acquisition, processing, and delivering real-time outputs. This design enhances the effectiveness of monitoring and improving sleep posture, particularly for elderly individuals who require special care. The workflow of the proposed system is illustrated in Fig. 1.

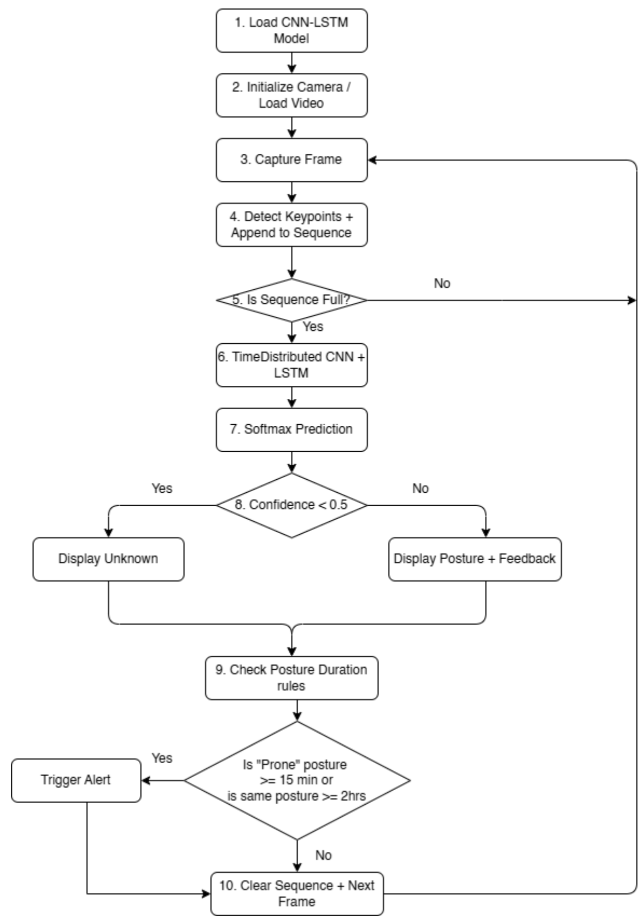

Fig. 1. Workflow diagram of the novel hybrid algorithm for vision-based sleep posture analysis integrating CNN, LSTM, and MediaPipe.

Sequence of the Sleep Posture Analysis System:

1) Load the pre-trained model, which consists of a TimeDistributed CNN for learning spatial features from keypoints and an LSTM for learning temporal patterns and continuity of sleep postures.

2) Activate the infrared (IR) camera to capture input data for posture detection. The camera must be positioned above the bed to clearly record the subject during sleep.

3) Capture video frames from the camera and apply MediaPipe to detect body landmarks (pose landmarks) such as shoulders, hips, and knees, which are essential for sleep posture analysis.

4) If keypoint detection is successful, append the current frame's set of keypoints to the accumulated keypoint sequence, preparing it for input into the sequence-based model.

5) Check the completeness of the accumulated keypoint sequence:

a) If incomplete, continue accumulating data.

b) If complete, feed the sequence into the CNN+LSTM model for behavior classification.

6) The output from the LSTM is passed through a Softmax Layer to obtain the probability values for each sleep posture, such as supine, left lateral, right lateral, and prone.

7) Evaluate the confidence level of the prediction:

a) If the confidence value is less than 50%, assign the label as “Unknown”.

b) If the confidence value is greater than 50%, display the detected posture along with feedback.

8) Display the name of the detected posture (e.g., supine) and optionally include a feedback alert if the posture is inappropriate (e.g., incorrect posture or risk of back pain).

9) Check whether the user has remained in the same posture for an extended period. Two primary conditions are monitored:

a) Staying in the prone position for more than 15 minutes.

b) Remaining in the same posture (any type) for more than 2 hours.

Clear the accumulated sequence of keypoints to prepare for the next cycle, and return to the frame-capturing stage for continuous real-time analysis.

C. Dataset Details

In this research, the dataset used for training and testing the continuous sleep posture detection and analysis system was collected from 20 participants (12 male, 8 female) aged between 19 and 21 years (mean = 20.1, SD = 0.8), with heights ranging from 1.55 m to 1.80 m and weights between 45 kg and 90 kg, resulting in a Body Mass Index (BMI) range of 18.7 to 29.4 kg/m². This variation in body types helps ensure the model's robustness across different physiques.

To simulate real-world conditions, participants wore a variety of sleep attire, including T-shirts with shorts, long-sleeved pajamas, and tank tops, and were recorded both with and without light blankets.

Video data were captured using a fixed infrared (IR) camera positioned directly above the bed at a height of 2.0 meters, simulating a typical bedside monitoring setup. Recordings were conducted in two primary lighting conditions: 1) Daytime: with natural and artificial ambient light (illuminance range: 100–300 lux), and 2) Nighttime: using only the camera's built-in IR illumination (illuminance: < 1 lux) to mimic a dark sleep environment.

The volunteers performed seven predefined posture transitions: supine to side (SUSI), side to supine (SISU), supine to prone (SUPR), prone to supine (PRSU), side to prone (SIPR), prone to side (PRSI), and side to side (SISI). Each participant carried out all seven posture transitions in the predetermined sequence, yielding a total of 280 video samples with 140 recorded under daytime lighting and 140 under nighttime lighting. The dataset was strategically partitioned into a training set (70 videos, 50%), a validation set (28 videos, 20%), and a test set (42 videos, 30%) for both lighting conditions. A stratified sampling method was rigorously applied during this partition to preserve the proportional distribution of the four target sleep posture classes within each subset. This approach prevents class imbalance in any data split and ensures that the model is evaluated on a representative sample of all postures, thereby increasing the reliability of the reported accuracy metrics and reducing the risk of skewed performance results.

D. Keypoints Extraction

The system begins by processing real-time video frames using the MediaPipe library, a machine learning framework developed by Google specifically for posture analysis. A pre-trained MediaPipe model was employed to detect 33 body landmarks in 3D space (X, Y, Z), where the Z-coordinate represents the depth relative to the camera.

The detected keypoints are converted into JSON format for each frame, storing landmark positions along with confidence scores. Keypoints include the head, shoulders, elbows, hands, hips, knees, and ankles. The keypoint extraction stage of the model is capable of processing up to 25 to 30 frames per second on standard hardware and remains effective even under partial occlusion of the body or in low-light environments.

The extracted keypoint data are then passed into the CNN and LSTM analysis stage to evaluate sleep postures and detect risky behaviors, such as prolonged prone sleeping. The use of MediaPipe in this process ensures high accuracy in landmark detection, with an average error margin of only ±5 pixels, making it highly suitable for real-world applications in nighttime sleep monitoring.

E. Pose Prediction

This system utilizes a deep learning hybrid approach that integrates CNN and LSTM to classify sleep postures from real-time video data. It is specifically designed to detect four primary sleep postures: supine, left lateral, right lateral, and prone. The model leverages the strengths of both architectures, with CNN responsible for extracting spatial features from body keypoints detected by MediaPipe in each frame, while LSTM analyzes temporal dependencies across consecutive frames.

The process begins with the use of a TimeDistributed layer in combination with a CNN structure, which is designed to process sequential data. This layer allows independent processing of each video frame while preserving the inter-frame relationships. For the LSTM component, a multi-layer architecture is employed to capture long-term dependencies in posture sequences, particularly for identifying risky behaviors such as remaining in the same posture for an extended duration or exhibiting abnormal posture transitions.

In the final stage, a SoftMax layer is applied to compute the probability distribution over each posture class, with the posture having the highest probability selected as the output. This design not only enhances the accuracy of posture classification but also enables the system to operate efficiently in an end-to-end manner.

The posture prediction model extracts keypoints from MediaPipe, passes them through a Convolutional Neural Network (CNN) to learn spatial features, and subsequently processes frame sequences with a Long Short-Term Memory

(LSTM) network to learn temporal relationships. The process can be expressed mathematically as follows:

1) Input video sequence: Let I denote a video sequence consisting of n frames:

$$I = \{F_1, F_2, \ldots, F_n\}, F_i \in \mathbb{R}^{H \times W \times 3}$$

where,

F_i is the i frame represented as a matrix with height H, width W, and 3 color channels (RGB).

$\mathbb{R}^{H \times W \times 3}$ represents the space of images of size H×W pixels with 3 color channels

2) Keypoint extraction with MediaPipe: Each frame F_i is processed by MediaPipe to extract 33 keypoints:

$$K_i = MediaPipe(F_i), K_i \in \mathbb{R}^{33 \times 3}$$

where,

K_i is the keypoint matrix for the i frame (33 keypoints × 3 coordinates: x, y, z).

Example keypoints

- Head ($K_1 = [x_1, y_1, z_1]$)
- Left Shoulder ($K_{11} = [x_{11}, y_{11}, z_{11}]$)
- Right Hip.($K_{24} = [x_{24}, y_{24}, z_{24}]$)

3) Spatial feature learning with CNN: Each keypoint matrix K_i is fed into a CNN to extract spatial features.

$$v_i = CNN(K_i), v_i \in \mathbb{R}^d$$

where,

v_i is the feature vector of length ($d = 64$)

CNN Architecture

The architecture consists of a 1D Convolutional layer (filter size 3) → ReLU → MaxPooling → Flatten.

Learned Patterns: The model learns patterns such as joint angles and movement speed of specific points.

4) Temporal sequence learning with LSTM: The sequence of spatial features $v = \{v_1, v_2, \ldots, v_n\}$ is analyzed by an LSTM network.

$$h = LSTM(v), h \in \mathbb{R}^m$$

where,

h is the temporal feature vector (m =32).

LSTM Mechanism:

- Forget Gate: Determines which information to discard or keep from the previous frame.
- Input Gate: Adds new information from the current frame.
- Output Gate: Outputs the relevant features.

5) Posture prediction:

$$\hat{y} = softmax(W_h + b), W \in \mathbb{R}^{c \times m}, b \in \mathbb{R}^c$$

where,

c is the number of postures (supine, left side-lying, right side-lying, prone).

$\hat{y}$ is the probability of each posture.

F. Alert Conditions

If the LSTM detects a continuous prone posture for more than 15 minutes (calculated from the number of frames and the frames per second rate), the system will send an alert signal. If the user remains in any single posture (whether supine, left lateral, right lateral, or prone) for a continuous period exceeding 2 hours, the system will issue an alert to prompt a change in position to prevent pressure ulcers.

The duration for which a posture is held is calculated from the number of consecutive frames and the video's frame rate. The formula to convert the frame count into minutes is as follows:

$$Duration_{min} = \frac{Frame\_count}{FPS \times 60}$$

where,

$Duration_{min}$ is the duration the posture is held (in minutes).

$Frame\_count$ is the number of consecutive frames in which the posture was detected.

FPS is the frame rate of the video (frames per second).

The sleep posture detection system was trained using sequence data over 100 epochs with Early Stopping employed to monitor categorical accuracy. Training was automatically halted when accuracy plateaued to prevent overfitting caused by excessive training.

Evaluation included generating a confusion matrix and calculating the accuracy score to analyze the model's ability to classify postures from videos not previously used in training.

The system predicts only the current (final) posture rather than the initial posture, as its primary purpose is to monitor user health by detecting potentially harmful current postures (such as prolonged prone sleeping.)

IV. RESULTS

A. Experimental Results

The experimental results of the sleep posture detection system under different lighting conditions indicated the performance metrics of the developed model separated by the four sleep posture classes: supine, left lateral, right lateral, and prone. The test dataset comprised 84 video clips recorded during daytime and nighttime. The postures with the largest number of samples were supine and prone (24 clips each), while left lateral and right lateral had equal samples (18 clips each).

From Table I, the test results indicated that the supine and prone postures achieved Precision, Recall, and F1-score values above 90%, reflecting that the model can accurately identify these two postures during daytime. The model's performance in detecting these postures under daylight conditions is visually represented in Fig. 2. In contrast, the classification performance for left lateral and right lateral postures slightly decreased, with Precision, Recall, and F1-score approximately ranging from 88% to 89%.

TABLE I. CLASS-WISE METRICS FOR DAYTIME IMAGES

Class	Test	Precision	Recall	F1-score
Supine position	12	93.3%	91.7%	92.5%
Left lateral decubitus position	9	89.1%	88.9%	89.0%
Right lateral decubitus position	9	88.5%	88.9%	88.7%
Prone position	12	91.0%	91.7%	91.3%

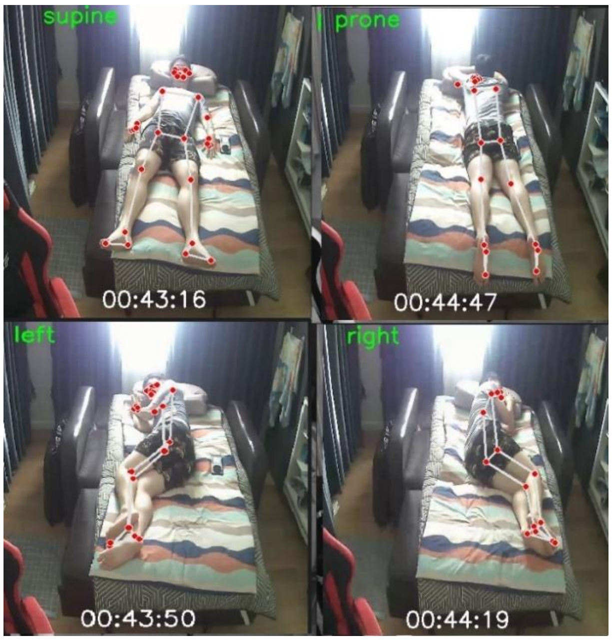

Fig. 2. Model testing of the hybrid algorithm for daytime images.

TABLE II. CLASS-WISE METRICS FOR NIGHTTIME IMAGES

Class	Test	Precision	Recall	F1-score
Supine position	12	90.9%	91.7%	91.3%
Left lateral decubitus position	9	85.7%	88.9%	87.3%
Right lateral decubitus position	9	84.6%	88.9%	86.7%
Prone position	12	88.9%	91.7%	90.3%

From Table II, the test results showed that the supine and prone postures achieved relatively high precision and recall values above 88%, with F1-scores around 90%, indicating that the model performed well in distinguishing clearly visible postures even under limited lighting conditions. Fig. 3 illustrates the hybrid algorithm model testing results for nighttime images. In contrast, the left and right lateral postures showed a slight drop in performance, with F1-scores approximately between 86% and 87%, reflecting the difficulty the model in distinguishing similar postures in low-light environments. Nonetheless, the overall results proved that the hybrid CNN, LSTM, and MediaPipe algorithm maintained strong classification capability even in low-light scenarios, which is a significant advantage for real-world nighttime applications.

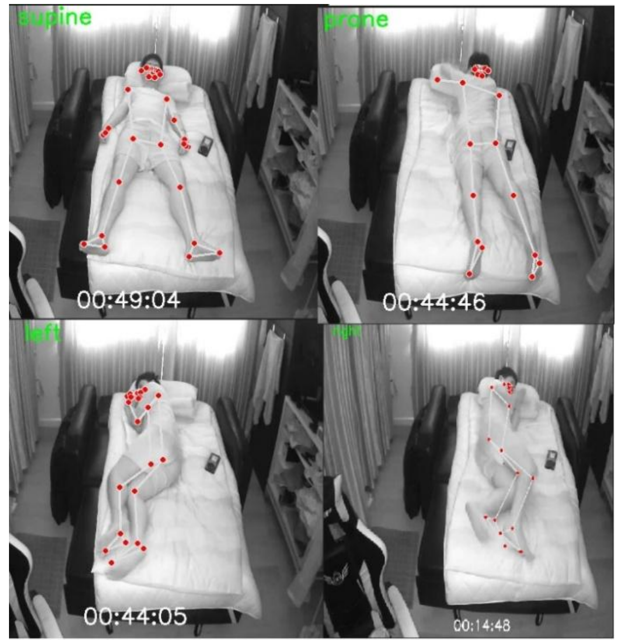

Fig. 3. Hybrid algorithm model testing for nighttime images.

Table III compares model performance in sleep posture analysis between daytime and nighttime images. The model achieved higher accuracy during daytime at 96.4%, while accuracy decreased to 92.8% at night, indicating that lighting conditions affect posture classification performance. The model tends to generate more misclassifications or confusion when processing low-light or nighttime images.

TABLE III. COMPARISON OF DAYTIME VS. NIGHTTIME ACCURACY

Lighting Condition	Video	Accuracy (%)
Daytime	42	96.4%
Nighttime	42	92.8%

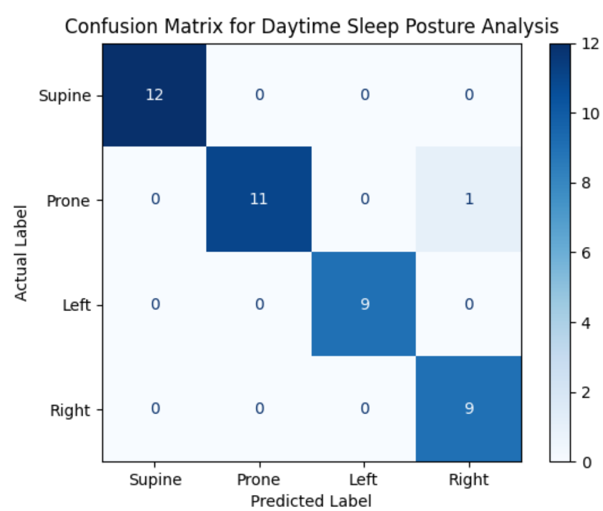

Fig. 4. Confusion matrix (daytime scenario).

From Fig. 4, the model's evaluation under daytime lighting conditions showed that it correctly classified all samples in the supine, left, and right classes with no mispredictions, demonstrating the model's strong ability to recognize distinctive features of these postures. However, there was one misclassification in the prone class, where one prone posture was predicted as right lateral. This confusion may have arisen due to the similarity of postures, such as a prone posture with the body or arms twisted in a manner resembling the lateral posture, leading to model misinterpretation.

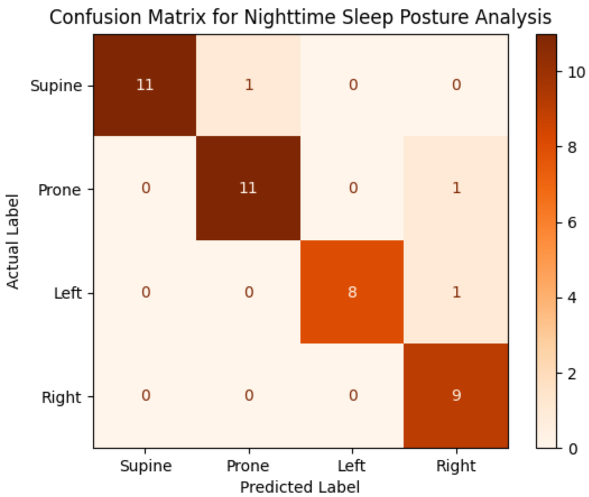

Fig. 5. Confusion matrix (nighttime scenario).

Fig. 5 presents the model's evaluation under a nighttime lighting scenario. Despite limited lighting, the model performed effectively with an overall accuracy of 92.8% based on 42 test clips comprising 12 supine clips, 12 prone clips, and 9 clips each for left and right lateral postures. Most posture classifications were correct, with only 3 misclassifications occurring due to confusion among similar postures. According to the confusion matrix, one supine posture was misclassified as prone, one prone posture was misclassified as right lateral, and one left lateral was misclassified as right lateral. The right lateral posture, however, was correctly classified in all nine instances. The confusion between prone, left, and right postures likely reflects the challenge of distinguishing postures under low light, where body features such as head tilt or arm positioning may be unclear. Additionally, partial occlusion by blankets reduces distinctive features needed to differentiate these postures.

V. Discussion

A. Comparative Analysis of Posture Detection Models

We conducted a performance comparison between the proposed hybrid CNN-LSTM algorithm and two related studies to demonstrate the advancements and strengths of our method.

The first related work, "Sleep Posture Classification using a Convolutional Neural Network" [14], presented a non-contact sleep posture monitoring system using depth sensor data and a CNN to classify main sleep postures, including supine, left lateral, right lateral, and prone. This system focused on real-world environments and generated vertical distance maps from depth images. It employed a multi-stream CNN architecture to enhance classification accuracy. Experimental results with 22 participants under three conditions (no blanket, thin blanket, thick quilt) showed average accuracies of 93.2%, 90.9%, and 88.6%, respectively. While this approach demonstrates robust performance, especially in handling blanket occlusions through depth information, it primarily relies on spatial data from a single frame. The absence of a temporal component, such as an LSTM, limits its ability to model posture transitions and continuously monitor posture duration for proactive health risk alerts – a key feature addressed by our proposed hybrid model.

The second study, "Sleep Posture Recognition Method Based on Sparse Body Pressure Features" [15], proposed a classification method using sparse pressure sensor data from an air mattress designed to reduce system cost and complexity. The air mattress had four pressure sensors located on the back, waist, hip, and leg, and utilized ensemble learning combining models such as SVM, XGBoost, RF, GP, and CNN to improve accuracy. This approach classified seven sleep postures (e.g., supine, side-lying or lateral, prone) with an accuracy of up to 95.63% and achieved 94.67% accuracy in classifying four posture types in real-world scenarios. The system also automatically adjusted mattress softness according to detected posture to enhance comfort. This sensor-based solution offers a valuable non-visual alternative; however, it requires specialized hardware (instrumented mattress), which can increase deployment cost and may not be as easily scalable or retrofittable as our vision-based approach. Furthermore, while effective for posture classification, the reliance on pressure maps may not capture the full skeletal alignment details that 3D keypoints from MediaPipe provide, and its ensemble model may lack the integrated spatio-temporal reasoning inherent in our end-to-end CNN-LSTM architecture.

TABLE IV. Comparative Experiments with Different Models

Model	Accuracy
MediaPipe + CNN + LSTM	96.40 % (Daytime)
depth sensor + CNN [14]	93.20 %
CNN + Autoencoder [15]	95.63 %

From Table IV, the accuracy comparison clearly showed the superiority of the proposed MediaPipe + CNN + LSTM model, achieving the highest accuracy of 96.4% compared to other models. This advantage results from the hybrid approach that integrates the strengths of multiple technologies: MediaPipe provides accurate real-time 3D pose information, CNN extracts spatial features from postures, and LSTM tracks temporal changes continuously. This comprehensive spatial-temporal capability enables the system to detect and classify complex postures more precisely. Furthermore, the model's robustness is reinforced by the diverse dataset, which includes variations in body types, clothing, and lighting, reducing the risk of overfitting to a specific demographic or environmental condition.

In contrast, the depth sensor + CNN model, with 93.2% accuracy, while effective with depth data, lacked critical temporal analysis. The CNN + Autoencoder model achieved 95.63%, but despite good dimensionality reduction, it cannot capture long-term posture relationships as effectively as LSTM.

This difference indicated that integrating these three technologies provided the system with superior detection and classification ability, particularly under real-world conditions involving lighting limitations and occlusions.

VI. Conclusion

Sleep posture significantly influences physical health and sleep quality, particularly for vulnerable populations such as the elderly and chronic patients. Traditional monitoring methods often involve invasive sensors or lack the capability for continuous, real-time analysis, limiting their practicality for long-term home care. To address this gap, this research developed a novel hybrid algorithm for non-invasive, real-time sleep posture analysis by effectively integrating MediaPipe, CNN, and LSTM technologies.

The system successfully detects and classifies four main sleep postures—supine, left lateral, right lateral, and prone—achieving a high accuracy of 96.4% under normal lighting and 92.8% under low-light conditions. It incorporates an automatic alert mechanism triggered when a prone posture is maintained for over 15 minutes or any single posture is held for more than 2 hours, addressing key health risks such as respiratory issues and pressure ulcers.

The key strength lies in the effective combination of MediaPipe's 3D keypoint detection, CNN's spatial feature extraction, and LSTM's temporal sequence analysis. This integration enables a comprehensive spatial-temporal understanding of sleep behavior, resulting in superior accuracy compared to depth-sensor or autoencoder-based approaches. The system maintains real-time performance with an overall end-to-end throughput of 22 frames per second on consumer-grade hardware, demonstrating strong practical feasibility for health monitoring applications.

For future work, we recommend: 1) expanding dataset diversity to include broader age groups and varied sleeping environments; 2) optimizing the model for mobile and IoT deployment; and 3) integrating with caregiver notification systems to enhance comprehensive health surveillance capabilities.

References

[1] Ravesloot, M. J. L., van Maanen, J. P., Dun, L., & de Vries, N. (2013). "The undervalued potential of positional therapy in positional obstructive sleep apnea syndrome." Sleep & Breathing, 17(1), 39–49. doi: 10.1007/s11325-012-0683-5.

[2] Gillespie, B. M., Walker, R. M., Latimer, S. L., Thalib, L., McInnes, E., & Chaboyer, W. P. (2021). "Repositioning for pressure injury prevention in adults." Cochrane Database of Systematic Reviews, (6), CD009958. doi: 10.1002/14651858.CD009958.pub3.

[3] A. Khan, C. Kim, S. Mey, K. S. Kim, E. Chung, and J. Lyu, "Ensemble deep learning for classifying sleeping posture of humans covered in blankets using RGB and thermal imaging," *IEEE Access*, vol. 12, pp. 169628–169642, Nov. 14, 2024, doi: 10.1109/ACCESS.2024.3498049.

[4] S. Robinson, A. M. Kumar, B. Janney, P. Sivakumar, et al., "Smart beds designed for monitoring and treatment of pressure ulcers," in Proc. 2023 IEEE 3rd Int. Conf. Technol., Eng., Manag. for Societal Impact Using Marketing, Entrepreneurship and Talent (TEMSMET), Feb. 2023, doi:10.1109/TEMSMET56707.2023.10150067.

[5] S. L. Bennett, R. Goubran, and F. Knoefel, "Long term monitoring of a pressure ulcer risk patient using thermal images," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 1461–1464, doi: 10.1109/EMBC.2017.8037110.

[6] R. Soleimani and D. Pesch, "Preventing Pressure Ulcers by Image based Posture Detection," in *Proc. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, L'Aquila, Italy, Jun. 22–24, 2023, doi: 10.1109/CBMS58004.2023.00186.

[7] Q. Hu, X. Tang, and W. Tang, "A real-time patient-specific sleeping posture recognition system using pressure sensitive conductive sheet and transfer learning," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6869–6879, Mar. 2021, doi: 10.1109/JSEN.2020.3043416.

[8] Q. Wan, H. Zhao, J. Li and P. Xu, "Human Sleeping Posture Recognition Based on Sleeping Pressure Image," in IEEE Sensors Journal, vol. 23, no. 4, pp. 4069-4077, 15 Feb.15, 2023, doi: 10.1109/JSEN.2022.3225290.

[9] S. Cheriyan, S. Sakthivel, T. Regula, K. K. SG, and S. Al Riyami, "Posture monitoring and classification for health care of bedridden patients using IoT and AI algorithms," in Proc. 2024 Int. Conf. Eng. Emerg. Technol. (ICEET), Bengaluru, India, 2024, pp. 1–6, doi: 10.1109/ICEET61466.2024.10569214

[10] D. Akinola, A. O. Oyedemi, and M. O. Ajinaja, "A Deep Learning-based Hybrid CNN-LSTM Model for Human Activity Recognition," Int. J. Sci. Res. Comp. Sci. Eng., vol. 12, no. 6, pp. 66–74, Dec. 2024.

[11] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," Procedia CIRP, vol. 99, pp. 650–655, 2021, doi: 10.1016/j.procir.2021.03.088.

[12] K. Tang, A. Kumar, M. Nadeem, and I. Maaz, "CNN-Based Smart Sleep Posture Recognition System," IoT, vol. 2, no. 1, pp. 119–139, 2021, doi: 10.3390/iot2010007.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, Dec. 2014. doi: 10.48550/arXiv.1412.3555.

[14] S. M. Mohammadi, M. Alnowami, S. Khan, D. -J. Dijk, A. Hilton and K. Wells, "Sleep Posture Classification using a Convolutional Neural Network," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 2018, pp. 1-4, doi: 10.1109/EMBC.2018.8513009.

[15] Li, Changyun, Guoxin Ren, and Zhibing Wang. 2025. "Sleep Posture Recognition Method Based on Sparse Body Pressure Features" Applied Sciences 15, no. 9: 4920. https://doi.org/10.3390/app15094920

[16] D. Cary, A. Jacques, and K. Briffa, "Correction: Examining relationships between sleep posture, waking spinal symptoms and quality of sleep: A cross sectional study," *PLOS ONE*, vol. 19, no. 7, p. e0306662, Jul. 2024, doi: 10.1371/journal.pone.0306662.