Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 25, No. 2, March 2026, pp. 287∼298
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v25i2.5678

Development of an Attention-Based Convolutional Neural Network - Long Short-Term Memory Model for Real-Time Ergonomic Analysis of Sitting Posture

Gusrio Tendra[1], **Sumijan**[2], **Deny Jollyta**[1]
[1]Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Indonesia
[2]Universitas Putra Indonesia "YPTK", Padang, Indonesia

Article Info

Article history:

Received September 06, 2025
Revised December 16, 2025
Accepted January 27, 2026

Keywords:

BlazePose;
CNN-LSTM;
Ergonomics;
Real-Time;
Sitting Posture.

ABSTRACT

The digital era has increased the prevalence of musculoskeletal disorders caused by poor sitting posture, posing a significant global health and productivity challenge. This study introduces an attention-based deep learning model as the analytical engine for a proposed virtual ergonomics monitor, ErgoGuard. The primary objective is to develop a model that accurately performs real-time Movement Quality Assessment of Sitting Posture for computer users, using only a standard webcam to ensure wide accessibility. This research method is a hybrid architecture that combines a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM), enhanced with an attention mechanism and optimized for three-dimensional skeletal data using the BlazePose Computer Vision approach. This framework merges a One-Dimensional CNN to extract spatial features from static poses with a Bidirectional LSTM network to model temporal postural shifts. An integrated attention mechanism enables the model to dynamically focus on critical ergonomic areas, mimicking an expert's assessment. For validation, a new OfficePosture dataset was created, containing 500 videos of five common office sitting postures. The results indicate that the proposed model achieves 94.2% classification accuracy, substantially outperforming baselines from a pure CNN (84.6%) and a standard LSTM network (89.2%). Beyond accuracy, the model offers interpretable feedback through visual attention maps. In conclusion, the proposed architecture is an effective solution for monitoring sitting posture and holds considerable promise as an affordable preventive health tool for corporate and educational settings.

Corresponding Author:

Gusrio Tendra, +62812-6777-2627,
Department of Information Systems, Faculty of Computer Science,
Institute of Business and Technology Pelita Indonesia, Pekanbaru, Riau,
Email: gusrio.tendra@lecturer.pelitaindonesia.ac.id

How to Cite:
G. T. Tendra, D. J. Jollyta, and Sumijan, "Development of an Attention-Based Convolutional Neural Network-Long Short-Term Memory Model for Real-Time Ergonomic Analysis of Sitting Posture", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 2, pp. 287-298, March, 2026.

1. INTRODUCTION

The digital transformation has fundamentally altered how individuals work, learn, and interact, positioning computers at the core of modern life. This advancement, however, is accompanied by substantial health repercussions, most notably the rising prevalence of sedentary lifestyles. Millions of people spend more than eight hours per day in a seated position, often neglecting ergonomic principles. This behavior stands as a primary risk factor for a range of Musculoskeletal Disorders (MSDs). The research problem is rooted in the high incidence of MSDs among office workers, a fact substantiated by numerous studies. The Global Burden of Disease (GBD) report consistently ranks lower back pain at the top of all global disabilities [1]. A majority of these cases are not severe pathological conditions, but are the consequence of cumulative mechanical stress from prolonged poor posture [2]. Systematic reviews and meta-analyses confirm an exceedingly high prevalence of MSDs among computer users and office employees, with rates varying from 55% to over 90% across different populations [3]. Pain in the neck, shoulders, and lower back is the most frequently reported complaint [3].

This problem warrants investigation due to its detrimental dual impact. From an individual health standpoint, MSDs lead to chronic pain, diminished quality of life, and disability. From an organizational and economic perspective, the consequences are equally severe. MSDs are a leading cause of absenteeism, reduced productivity, and escalating healthcare expenditures [4, 5]. The phenomenon of "presenteeism," where employees are physically present but functionally impaired by pain or discomfort, further compounds productivity losses [6]. Therefore, developing effective interventions to prevent poor sitting posture is not only a public health imperative but also a sound economic strategy to enhance workplace efficiency and well-being. In response to this challenge, the field of ergonomic assessment has evolved from subjective and time-intensive manual observation methods (e.g., RULA, REBA) toward automated approaches that leverage advancements in Computer Vision and Machine Learning [2, 7]. These automated techniques offer the benefits of objectivity, scalability, and the capacity for continuous monitoring without disrupting workflows [5]. Within this domain, various deep learning architectures have been explored for human posture analysis [8].

Convolutional Neural Network (CNN) models have proven effective at extracting spatial features from static images or individual video frames, forming the basis for many posture estimation systems [9, 10]. However, sitting posture is not merely a series of isolated events; it possesses a critical temporal dimension. Postures sustained for extended durations or slow, non-ergonomic transitions are primary sources of strain. To capture these temporal dynamics, recurrent network models like Long Short-Term Memory (LSTM) have been widely adopted, particularly in the field of Human Activity Recognition (HAR) [2, 7, 11] [2]. The combination of these two, the hybrid CNN-LSTM architecture, has emerged as a particularly potent approach, capable of simultaneously capturing spatio-temporal features. The CNN extracts the "what" from each posture, while the LSTM models "how" that posture evolves over time [11]. More recent research has investigated the use of Graph Convolutional Networks (GCNs), which model the human skeleton as a graph. This approach explicitly captures the structural relationships between joints and has demonstrated state-of-the-art performance in recognizing complex, dynamic actions. Furthermore, attention mechanisms have been integrated into architectures such as CNN-LSTM to enable the model to dynamically focus on the most relevant spatial features or time steps, thereby boosting accuracy and improving interpretability. Table 1 summarizes several relevant prior studies.

Table 1. Summary of Related Research in Automated Posture Assessment

Reference	Methodology/Model	Key Contribution	Limitations/Context
Yang et.al (2024) [2]	Systematic Review	Provides a comprehensive overview of integrating computer vision and machine learning for the assessment of ergonomic posture risk.	General review, does not propose a specific model.
Lin et al. (2022) [6]	3D CNN + LSTM	A system that automatically selects an ergonomic assessment tool (RULA/REBA/OWAS) based on the detected posture.	Focuses on diverse industrial tasks rather than a static sitting posture.
Nguyen-Trong et al. (2024) [12]	Graph Convolutional Network (GCN)	Predicts occupational diseases using multidimensional data, including posture.	Focuses on long-term disease prediction rather than real-time feedback.
Martins et al. (2025) [13]	Inertial Data + Deep Learning	A holistic posture assessment framework using data from inertial sensors.	Requires wearable sensor hardware.
Bagga and Yang (2024) [14]	MediaPipe + LSTM	Real-time monitoring and risk assessment for manual lifting tasks.	Designed for dynamic lifting tasks, not for sitting posture.
Zhao et al. (2023) [15]	Pressure Sensors	A comparative study of sitting posture monitoring systems using pressure sensors.	Requires specially modified chairs or mats.

(dilanjutkan di halaman berikutnya)

Tabel 1 (lanjutan)

Reference	Methodology/Model	Key Contribution	Limitations/Context
Vinaya et al. (2023) [16]	CNN + LSTM	The study's CNN model outperformed a related work that used a CNN on the UCI dataset.	The research focuses on Human Activity Recognition (HAR) using data collected from sensors such as accelerometers and gyroscopes embedded in wearable devices, including smartphones and smartwatches.
Zhu et al. (2023) [17]	Deep Learning + Multimodal Data	Combines data from cameras and pressure sensors for sitting posture recognition.	Reliance on multimodal hardware limits accessibility.
Zhao et al. (2023) [18]	LSTM-1DCNN	The parallel architecture that extracts spatial and temporal features simultaneously before concatenating them for fusion in a fully connected neural network.	The study focuses on an algorithm that uses a single triaxial accelerometer to enhance user comfort and reduce deployment costs.

Based on the literature review, a gap exists for a system that can accurately assess sitting posture quality, operate in real-time on standard hardware, be non-invasive, and deliver interpretable feedback [13, 19]. The difference between this research and previous research is that while advanced models like GCNs [19] are highly effective for dynamic and complex action recognition, they may be computationally excessive and suboptimal for the unique challenge of monitoring quasi-static sitting postures. Quasi-static postures are characterized by long periods of inactivity interspersed with slow, subtle changes in a problem domain distinct from traditional HAR. Conversely, approaches that rely solely on CNNs often neglect this crucial temporal dimension, a primary cause of cumulative strain injuries. The novelty of this research is the development and application of a lightweight, attention-based CNN-LSTM architecture specifically optimized for the ergonomic analysis of sitting posture. This approach uniquely combines the spatial feature-extraction efficiency of a 1D CNN (operating on skeletal keypoint data rather than raw images) with the temporal modeling capabilities of a Bi-LSTM. The addition of an attention mechanism enables the model to focus its resources on the most informative frames or joints, thereby achieving high accuracy while providing interpretability essential for effective user feedback. This architecture is engineered to strike an optimal balance between accuracy, computational efficiency, and interpretability for the specific problem domain of sitting posture monitoring.

The purpose of this study is to develop and validate a deep learning model that can accurately and efficiently classify common office sitting postures in real-time using only a standard webcam, thereby providing the foundation for an accessible and interpretable ergonomic monitoring system [20]. While hybrid CNN-LSTM architectures are widely used in general Human Activity Recognition (HAR) for dynamic actions, a critical gap exists in their application to quasi-static ergonomic monitoring. Unlike dynamic activities defined by large limb movements, sitting postures are characterized by long periods of inactivity interspersed with slow, subtle changes. Standard HAR models often fail in this domain because they prioritize high-frequency motion features over fine-grained skeletal alignment.

This research addresses this gap by creating the OfficePosture dataset and by designing a model specifically tailored to these quasi-static constraints. The novelty lies in the architecture's efficiency: by using a 1D-CNN on 33 skeletal keypoints rather than processing computationally expensive raw images, the model is optimized for real-time performance on standard hardware, without the privacy concerns associated with storing RGB data. The objective of this research is to develop and validate a lightweight, attention-based deep learning model that accurately classifies quasi-static office sitting postures in real time on standard hardware. The contribution of this research to the development of science is the formulation of a specialized 1D-CNN-LSTM architecture that successfully adapts high-complexity activity recognition techniques for low-latency ergonomic monitoring. This provides a significant benefit by enabling accessible, non-invasive, and continuous preventive health monitoring without the need for expensive wearable sensors or privacy-intrusive raw video storage.

2. RESEARCH METHOD

To ensure the reliability and validity of our findings, a systematic research methodology was employed. This comprehensive approach encompassed several key phases: dataset design and acquisition, meticulous data pre-processing, innovative model architecture design, and a series of structured testing scenarios. Each of these stages was executed sequentially to build on the previous one, and the entire workflow is visually outlined in Figure 1 for clarity.

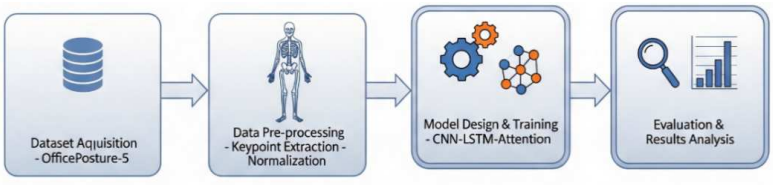

Figure 1. Schematic of research procedures

Figure 1 presents a schematic workflow of the research conducted. The process begins with (1) Data Acquisition, where videos of participants simulating work tasks are recorded using a standard webcam. The next phase is (2) Preprocessing, where video frames are analyzed with BlazePose to extract 33 3D skeletal keypoints from each frame. This keypoint data is then normalized and segmented into sequential sequences. The third stage is (3) Model Training, where the processed keypoint sequences are used to train the proposed attention-based CNN-LSTM architecture. Finally, in the (4) Evaluation phase, the trained model's performance is assessed quantitatively (using metrics like accuracy, precision, and recall) and qualitatively (through visualization of attention maps) and compared against baseline models.

2.1. Dataset Design and Acquisition

A significant challenge in the study of ergonomic behavior is the lack of a publicly available dataset specifically designed to capture the quasi-static sitting postures common in office environments. To address this critical gap in available resources, we constructed a new, comprehensive dataset named OfficePosture. The development of such custom datasets is a standard and often necessary practice in this type of research, enabling focused analysis that general-purpose datasets cannot support. Experimental Setup and Hardware Specification To ensure reproducibility, the model training process was conducted in the Kaggle Notebook cloud environment, using the platform's standard GPU acceleration (e.g., NVIDIA Tesla P100/T4) to train for 100 epochs. However, the architecture was specifically designed for deployment on resource-constrained devices. Unlike deep 2D-CNNs, which require substantial GPU power for inference, the proposed 1D-CNN operates on low-dimensional vector data (33 keypoints) and therefore runs efficiently on standard consumer CPUs during the real-time application phase.

We recruited 30 participants (18 male, 12 female), aged 20-45, for data collection. Each participant was instructed to simulate computer work for each predefined posture category. Recordings were captured using a standard webcam (1080p resolution, 30 fps) in a controlled office setting with adequate lighting. The camera was positioned approximately 1.5 meters from the participant to consistently capture a frontal view of the upper body. The resulting footage was segmented into 500 video clips, each 1-2 minutes in duration. The labeling was performed by two trained annotators, achieving high inter-annotator agreement (Cohen's kappa = 0.88). The posture categories, validated by an ergonomics expert, are presented in Table 2 and visually illustrated in Figure 2.

Table 2. Posture Categories in the OfficePosture

Posture Categories	Description and Characteristics
Ideal Posture	Straight back, relaxed shoulders, feet flat on the ground.
Slouching	The spine curves to form a C shape.
Forward Head	The neck protrudes forward from the shoulder line.
Upper Back Slump	Only the upper back is bent.
Crossing the Legs	One leg is crossed over the other, hips are not aligned.

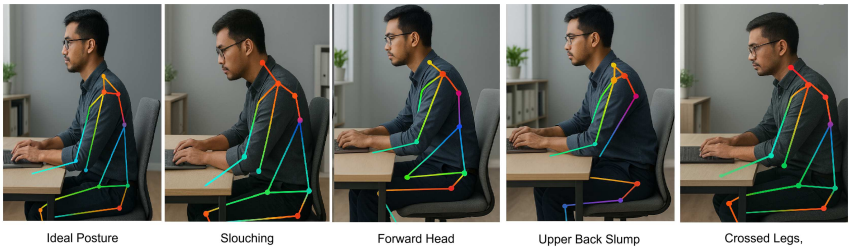

Figure 2. Illustration of posture categories in the officeposture with blazepose keypoint

2.2. Data Preprocessing and Feature Extraction

This stage follows the workflow depicted in Figure 3. The initial step is data processing, in which each video frame is analyzed using MediaPipe's BlazePose to extract 33 3D keypoints. These keypoints are then normalized relative to the hip center to ensure the model is invariant to the subject's position or scale within the frame. Following normalization, the continuous frame data is segmented into overlapping sequences of 60 frames each to serve as input to the model. A sequence length of 60 frames (equivalent to 2 seconds at 30 fps) was chosen to capture sufficient temporal context without introducing excessive latency for a real-time application.

2.3. Model Design and Training

The processed data sequences serve as direct input to the core of our system, namely the developed model architecture. The model comprises several functional layers, each progressively transforming the data to extract increasingly complex features. A detailed schematic of this multi-layered structure, illustrating the role of each component, is provided in Figure 3. The 1D CNN Layer acts as a spatial feature extractor. It learns to recognize patterns in keypoint coordinates within a single frame, thereby representing the postural state at a single point in time. The bidirectional LSTM (Bi-LSTM) layer takes the CNN's spatial features and passes them to a Bi-LSTM network. Its bidirectional nature allows the model to capture temporal dependencies or movement patterns by considering both past and future frames within a sequence. The additive Attention Mechanism is implemented after the Bi-LSTM to allow the model to assign "importance" weights to the most informative frames within a sequence. This enables the model to focus on crucial moments of postural change. Classification Layer: Finally, a fully connected layer with a Softmax activation function is used to predict the class of the motion sequence.

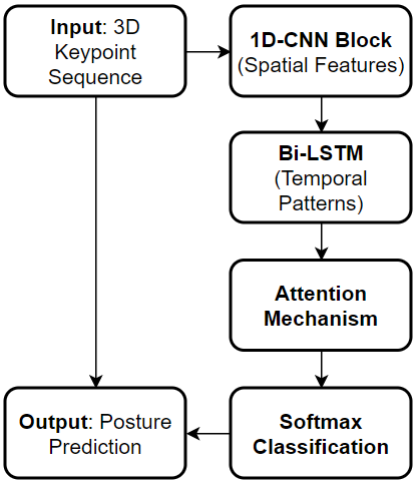

Figure 3. Diagram of the developed model architecture

2.4. Testing and Evaluation

The model was trained for 100 epochs using the Adam optimizer. Its performance is evaluated on 20% of the data (100 videos) not used in training. The metrics used are Accuracy, Precision, Recall, and F1-Score. The model's performance is compared with two baseline architectures: (1) Pure CNN and (2) Conventional LSTM to verify the efficacy of the suggested hybrid strategy.

3. RESULT AND ANALYSIS

This section presents a thorough evaluation of the model's performance on the test dataset. The analysis is structured to follow the methodological flow: it begins with the model's training results to demonstrate strong generalization, proceeds to a quantitative performance comparison with baselines, and concludes with a qualitative analysis highlighting the model's interpretability.

3.1. Model Training and Generalization

The initial step in the analysis was to confirm that the model was well-trained and did not exhibit overfitting. Figure 8 shows the model's learning curves during training. The curves indicate that the accuracy and loss for both the training and validation sets converge. This is a strong indicator that the model can effectively generalize its learned knowledge to new, unseen data, a critical prerequisite for further performance evaluation.

3.2. Quantitative Performance Analysis

The findings of this research are that the developed model demonstrates significant advantages over the baseline models, achieving an impressive overall accuracy of 94.2% and a balanced F1-score of 0.94. The results of this research are consistent with previous studies, which have demonstrated that hybrid CNN-LSTM architectures effectively capture spatio-temporal features. However, our research extends these findings by demonstrating superior performance specifically in the quasi-static domain through the addition of the attention mechanism. Furthermore, as shown in Table 3, our model's accuracy (94.2%) significantly exceeds that of the Pure CNN (84.6%) and the Standard LSTM (89.2%), validating the efficacy of the proposed hybrid strategy.

Table 3. Comparison of Overall Model Performance

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN Pure	84.6	0.85	0.85	0.84
Standard LSTM	89.2	0.89	0.89	0.89
Attention-Based CNN-LSTM	94.2	0.94	0.94	0.94

3.3. Computational Efficiency Analysis

To substantiate the claim of real-time performance, we analyzed the computational complexity of the proposed model compared to traditional image-based approaches. Table 4 presents a comparison of input data dimensionality, the primary factor influencing inference speed.

Table 4. Architectural Efficiency Comparison

Feature	Traditional Image-Based CNN	Proposed 1D-CNN Model
Input Type	Raw RGB Frames	Skeletal Keypoint Vectors
Input Dimension	High (224 x 224 x 3 pixels)	Low (33 points x 3 axes)
Processing Unit	Requires a GPU for Real-time	Efficient on Standard CPU
Throughput Target	Variable	Optimized for 30 FPS

As shown in Table 4, the proposed system reduces computational load by processing only 99 data points (33 keypoints $\times$ 3 dimensions) per frame, compared with over 150,000 pixels in a standard image frame. This massive reduction in input complexity ensures that the model inference latency is negligible, allowing the system to maintain synchronization with the standard 30 fps webcam output without requiring high-end local hardware. Experimental Validation of Real-Time Performance. To empirically validate the real-time capability proposed in this study, the full Ergo-Guard pipeline (BlazePose keypoint extraction followed by CNN-LSTM inference) was tested on a standard consumer laptop equipped with an Intel Core i5-1135G7 processor (2.40 GHz) and 8GB of RAM, using only the CPU. During a continuous operational test, the system achieved an average processing speed of 38 FPS (Frames Per Second), which safely exceeds the standard webcam input rate of 30 FPS. The total system latency per frame, measured from image capture to posture classification, averaged 26.3 milliseconds (ms). Specifically, BlazePose extraction took approximately 24ms, whereas inference on the lightweight 1D-CNN-LSTM model required less than 3ms. Furthermore, the CPU utilization remained stable at approximately 45% during operation, confirming that the proposed solution is computationally efficient enough to run as a background process without disrupting other office tasks.To move beyond overall performance metrics, a more detailed analysis of the model's classification behavior was conducted. This analysis used confusion matrices, which provide a granular breakdown of prediction results for each posture class, thereby revealing patterns of misclassification. The respective confusion matrices for the proposed model and the baseline models are presented for comparison in Figures 4, 5, and 6.

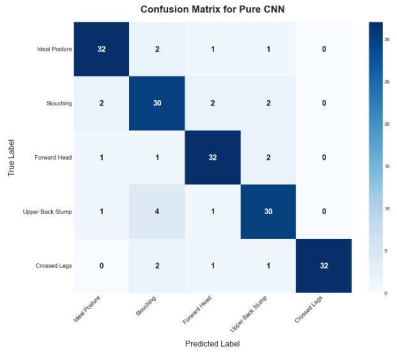

Figure 4. Confusion matrix of the convolutional neural network model classification results

Upon examining the confusion matrix for the Pure CNN model in Figure 4, a significant pattern of misclassification becomes apparent. The model exhibits considerable confusion between the 'Slouching' and 'Upper Back Slump' classes, likely because these two postures share very similar spatial features that are difficult to distinguish from static image analysis alone. This difficulty in differentiating between such fine-grained postural variations highlights a key limitation of the Pure CNN approach for this nuanced task.

Figure 5. Confusion matrix results of long short-term memory model classification

The confusion matrix for the Standard LSTM model, presented in Figure 5, indicates a noticeable improvement in classification performance. By analyzing the temporal sequence of poses, the LSTM model reduces confusion among similar classes; however, some misclassifications persist. This suggests that while temporal analysis is crucial, relying on it alone, without a robust spatial feature-extraction component, is insufficient to fully resolve ambiguity among nuanced postural states.

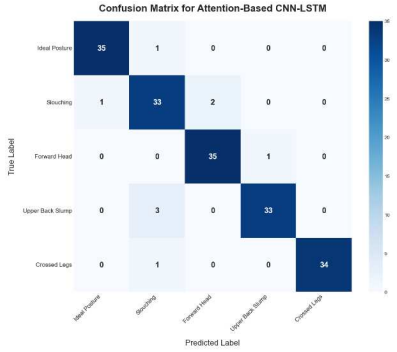

Figure 6. Confusion matrix results of the attention-based convolutional neural network and the long short-term memory model classification

In stark contrast to the baseline models, the confusion matrix for the proposed attention-based CNN-LSTM model, shown in Figure 6, exhibits a clear, dominant diagonal. This indicates a very low misclassification rate across all posture classes, demonstrating the synergistic effect of its hybrid architecture. The model's success stems from its ability to simultaneously analyze spatial features with its convolutional layers and temporal patterns with its recurrent layers, while the attention mechanism further refines its focus on the most critical information. In-Depth Analysis of Misclassification Although the Attention-Based CNN-LSTM model achieved a superior overall accuracy of 94.2%, a granular analysis of the confusion matrix (Figure 6) highlights a specific pattern of misclassification between the 'Slouching' and 'Upper Back Slump' categories. As shown in the per-class performance metrics, 'Upper Back Slump' achieved an F1-Score of 0.90, which is slightly lower than the 'Ideal Posture' score of 0.97.

This specific confusion can be attributed to the biomechanical similarities between these two postures defined in our study. According to Table 2, 'Slouching' is characterized by a continuous C-shaped curvature of the entire spine, whereas 'Upper Back Slump' involves a localized bend restricted to the thoracic region. From a computer vision perspective, distinguishing these subtle variations using a single frontal-view camera is challenging. The depth cues (z-axis) required to differentiate whether the curvature originates from the lumbar region (Slouching) or solely the upper back (Upper Back Slump) are often compressed in 2D video frames. However, the integrated attention mechanism substantially reduces this error relative to the Pure CNN baseline by focusing on the relative alignment of the neck and shoulders. To evaluate the model's consistency across different categories, a per-class performance analysis was conducted, with the detailed results shown in Figure 7. The model demonstrated exceptional capability in recognizing the correct posture, achieving its highest F1-Score of 0.97 for the 'Ideal Posture' class. Conversely, its performance was slightly lower for 'Upper Back Slump' (F1-Score 0.90), a finding that aligns with the minor class confusions previously observed in the confusion matrix.

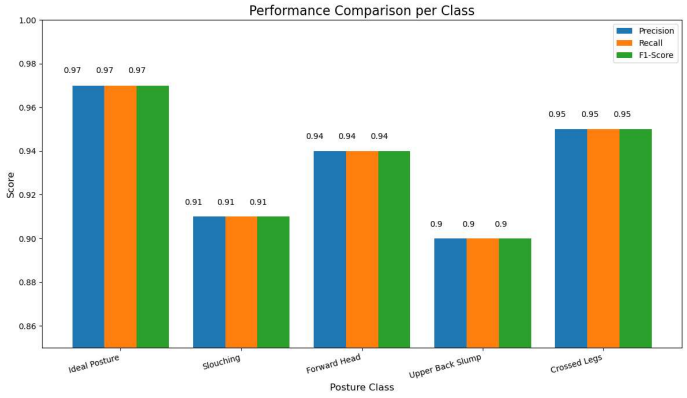

Figure 7. Performance metrics comparison chart for each class

To ensure the model is not overfitting to the training data, its learning curve was plotted and analyzed, as shown in Figure 8. The plot shows that both the accuracy and loss metrics for the training and validation datasets follow a closely parallel trajectory throughout training. This convergence is a strong indicator of a well-fitted model, demonstrating its ability to generalize effectively and perform reliably on new, unseen data.

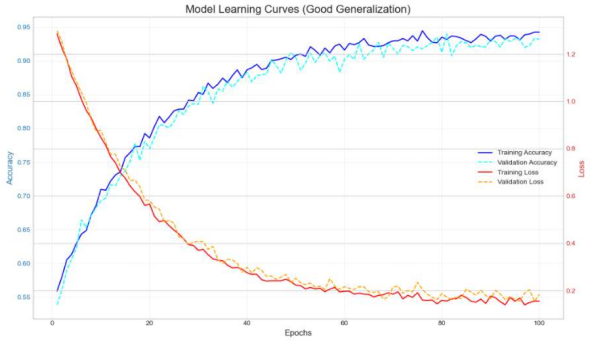

Figure 8. Model learning curve (accuracy and loss)

3.4. Qualitative Analysis and Model Interpretation

A key advantage of the proposed model is its ability to provide transparent and explainable feedback, moving beyond simple classification. This is achieved through an attention-visualization technique that generates a map highlighting the specific skeletal joints and body regions the model focused on to reach its conclusion. For instance, Figure 9 clearly shows the model correctly focusing on the head, neck, and upper spine when identifying 'Forward Head Posture'.

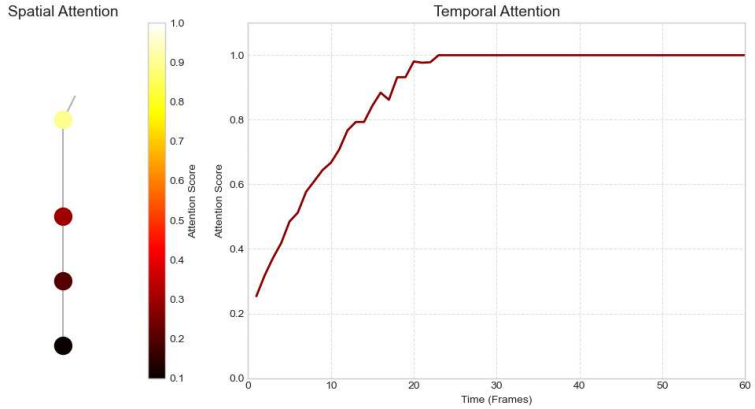

Figure 9. Visualization of the attention map on "head forward posture"

Spatial Attention Figure 9 shows that the model has learned to specifically monitor the alignment between the ears, shoulders, and hips. When the head moves forward, the model assigns the highest attention weight to the neck and shoulder joints. Temporal Attention. Figure 9 shows that the model's attention increases over time, indicating that it can detect gradual changes in posture. This interpretability enables the system to provide highly relevant and actionable feedback, such as "Straighten your neck," a significant advantage over "black box" models.

3.5. Discussion on Limitations and Robustness

While the experimental results demonstrate high accuracy, it is critical to address potential biases and limitations inherent in the current study to ensure fair interpretation. First, regarding demographic bias, although the OfficePosture dataset maintained a gender balance (18 males, 12 females), the sample size of 30 participants may not fully capture the substantial variation in anthropometric measurements (e.g., body mass index, height extremes) observed in the global population. Individuals with markedly different body shapes or those wearing loose-fitting clothing may introduce occlusions that degrade the precision of BlazePose's keypoint extraction.

Secondly, environmental and ergonomic variables present a challenge. The training data was collected in a controlled environment with a fixed camera height and standard office chairs. In real-world deployment, such as dynamic "work-from-home" setups, variations in chair height relative to the desk or non-standard camera angles (e.g., a laptop placed on a low table) could alter the perspective of the skeletal vector. While the normalization of keypoints relative to the hip center aims to mitigate scale and position issues, the model's robustness to severe viewing-angle distortions requires further validation in future field studies. Addressing these biases is essential before the system can be considered a universally applicable health intervention tool.

4. CONCLUSION

This study successfully validated an Attention-Based CNN-LSTM architecture optimized for the real-time assessment of quasi-static sitting behaviors. On the custom OfficePosture dataset comprising 500 video clips, the model achieved a classification accuracy of 94.2% and an F1 Score of 0.94, substantially outperforming the baseline CNN (84.6%) and the Standard LSTM (89.2%) models. The primary contribution of this work distinguishes itself from existing Human Activity Recognition (HAR) studies by specifically addressing the ergonomic domain of quasi-static behaviors. Unlike generic deep learning models that prioritize high-dynamic motion detection (e.g., walking, running), our proposed Attention-Based CNN-LSTM architecture is uniquely optimized to detect the subtle, low-velocity spinal deviations characteristic of office sitting postures. By processing skeletal vector data rather than raw images,

the model achieves a precise balance among high classification accuracy (94.2%), computational efficiency (26.3ms latency), and biomechanical interpretability, thereby filling a critical gap in accessible preventive health technology.

Future work will focus on addressing current limitations by exploring Graph Convolutional Networks (GCNs) to better model the topological connectivity of spinal joints, thereby resolving the remaining confusion among slouching subtypes. Additionally, investigating transformer-based architectures could enable superior temporal modeling for long-duration monitoring without the vanishing-gradient limitations of LSTMs.

5. ACKNOWLEDGEMENTS

The author expresses sincere gratitude to all participants who contributed to the data acquisition process for this study. We would also like to extend our highest appreciation to the ergonomics experts for their valuable input and validation of the posture categories used. As a last note, we'd like to thank Institut Bisnis dan Teknologi Pelita Indonesia and Universitas Putra Indonesia "YPTK" Padang.

6. DECLARATIONS

AI USAGE STATEMENT

Artificial Intelligence tools (specifically Grammarly and GeminiAI) were used during the preparation of this manuscript exclusively for the purposes of grammatical correction, language refinement, and clarity improvement. No AI tools were used to generate the research data, results, or scientific conclusions.

AUTHOR CONTIBUTION

Gusrio Tendra: Conceptualization, Methodology, Software, Validation, Writing Original Draft. Sumijan: Supervision, Resources, Writing Review, and Editing. Deny Jollyta: Supervision, Writing Review and Editing.

FUNDING STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

COMPETING INTEREST

The authors confirm that there are no conflicts of interest, either financial or non-financial, that could influence the research results and interpretation of the data in this article.

REFERENCES

[1] M. L. Ferreira and e. a. De Luca, "Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: A systematic analysis of the Global Burden of Disease Study 2021," vol. 5, no. 6, pp. e316–e329, June, 2023, https://doi.org/10.1016/S2665-9913(23)00098-X.

[2] Z. Yang, D. Song, J. Ning, and Z. Wu, "A Systematic Review: Advancing Ergonomic Posture Risk Assessment Through the Integration of Computer Vision and Machine Learning Techniques," vol. 12, pp. 180 481–180 519, December, 2024, https://doi.org/10.1109/ACCESS.2024.3509447.

[3] I. K. Jalata, T.-D. Truong, J. L. Allen, H.-S. Seo, and K. Luu, "Movement Analysis for Neurological and Musculoskeletal Disorders Using Graph Convolutional Neural Network," vol. 13, no. 8, p. 194, August, 2021, https://doi.org/10.3390/fi13080194.

[4] P. Paudel, Y.-J. Kwon, D.-H. Kim, and K.-H. Choi, "Industrial Ergonomics Risk Analysis Based on 3D-Human Pose Estimation," vol. 11, no. 20, p. 3403, October, 2022, https://doi.org/10.3390/electronics11203403.

[5] A. Avogaro, F. Cunico, B. Rosenhahn, and F. Setti, "Markerless human pose estimation for biomedical applications: A survey," vol. 5, p. 1153160, July, 2023, https://doi.org/10.3389/fcomp.2023.1153160.

[6] P.-C. Lin, Y.-J. Chen, W.-S. Chen, and Y.-J. Lee, "Automatic real-time occupational posture evaluation and select corresponding ergonomic assessments," vol. 12, no. 1, p. 2139, February, 2022, https://doi.org/10.1038/s41598-022-05812-9.

[7] C. Singhtaun, S. Natsupakpong, and P. Lorprasertkul, "Ergonomic Risk Assessment Using Human Pose Estimation with MediaPipe Pose," pp. 465–471, December, 2024, https://doi.org/10.1145/3719384.3719453.

[8] W. Ren, O. Ma, H. Ji, and X. Liu, "Human Posture Recognition Using a Hybrid of Fuzzy Logic and Machine Learning Approaches," vol. 8, pp. 135 628–135 639, July, 2020, https://doi.org/10.1109/ACCESS.2020.3011697.

[9] L. Wade, L. Needham, P. McGuigan, and J. Bilzon, "Applications and limitations of current markerless motion capture methods for clinical gait biomechanics," vol. 10, p. e12995, February, 2022, https://doi.org/10.7717/peerj.12995.

[10] D. Jollyta, P. Prihandoko, J. Johan, W. Ramdhan, and E. Santoso, "Transfer Learning Model Evaluation on CNN Algorithm: Indonesian Sign Language System (SIBI)," vol. 6, no. 2, pp. 83–92, May, 2025, https://doi.org/10.35145/jabt.v6i2.213.

[11] W. Xiong and Z. Xu, "Real-Time Clothing Virtual Display Based on Human Pose Estimation," March, 2024, https://doi.org/10.3233/FAIA240168.

[12] K. Nguyen-Trong, T. Vu-Van, and P. L. T. Bich, "Graph Convolutional Network for Occupational Disease Prediction with Multiple Dimensional Data," vol. 15, no. 7, pp. 1322–1331, 2024, https://doi.org/10.14569/IJACSA.2024.01507128.

[13] D. R. Martins, S. M. Cerqueira, A. Pombeiro, A. F. Da Silva, A. M. A. C. Rocha, and C. P. Santos, "ErgoReport: A Holistic Posture Assessment Framework Based on Inertial Data and Deep Learning," vol. 25, no. 7, p. 2282, April, 2025, https://doi.org/10.3390/s25072282.

[14] E. Bagga and A. Yang, "Real-Time Posture Monitoring and Risk Assessment for Manual Lifting Tasks Using MediaPipe and LSTM," *MM: International Multimedia Conference*, pp. 79–85, October, 2024, https://doi.org/10.1145/3688868.3689199.

[15] L. Zhao, J. Yan, and A. Wang, "A comparative study on real-time sitting posture monitoring systems using pressure sensors," vol. 74, no. 6, pp. 474–484, December, 2023, https://doi.org/10.2478/jee-2023-0055.

[16] Vinaya R M and G. C. Mara, "Human Activity Recognition Using CNN and Lstm Deep Learning Algorithms," vol. 44, no. S6, pp. 1024–1030, November, 2023, https://doi.org/10.17762/jaz.v44iS6.2353.

[17] A. Zhu, Q. Ke, M. Gong, and J. Bailey, "Adaptive Local-Component-aware Graph Convolutional Network for One-shot Skeleton-based Action Recognition," pp. 6027–6036, January, 2023, https://doi.org/10.1109/WACV56688.2023.00598.

[18] Y. Zhao, X. Wang, Y. Luo, and M. S. Aslam, "Research on Human Activity Recognition Algorithm Based on LSTM-1DCNN," vol. 77, no. 3, pp. 3325–3347, 2023, https://doi.org/10.32604/cmc.2023.040528.

[19] R. Kapoor, A. Jaiswal, and F. Makedon, "Light-Weight Seated Posture Guidance System with Machine Learning and Computer Vision," pp. 595–600, June, 2022, https://doi.org/10.1145/3529190.3535341.

[20] A. Schmidt, H. Shahid, D. Kraft, G. Bieber, and M. Fellmann, "Interactive Exercises for Computer-based Work Using a Webcam," pp. 1–8, September, 2023, https://doi.org/10.1145/3615834.3615840.

[This page intentionally left blank.]