Jurnal Testing dan Implementasi Sistem Informasi
Vol 3(2) 2025 : 94-107

HAND POSE CLASSIFICATION USING MEDIAPIPE HANDS AND CNN-LSTM FOR AUGMENTED REALITY BASED INTRAVENOUS INFUSION LEARNING

Yenny Desnelita[1*], Muhammad Siddik[2], Lita[3], Alyauma Hajjah[4], Gustientiedina[5]
Department Information Systems, Institut Bisnis dan Teknologi Pelita Indonesia, Indonesia[1]
Departement Informatics Engineering, Institut Bisnis dan Teknologi Pelita Indonesia, Indonesia[2,4,5]
Departement Nursing Science, Universitas Hangtuah Pekanbaru, Indonesia[3]
yenny.desnelita@lecturer.pelitaindonesia.ac.id[1*], siddik@lecturer.pelitaindonesia.ac.id[2], lita@htp.ac.id[3], alyauma.hajjah@lecturer.pelitaindonesia.ac.id[4], gustientiedina@lecturer.pelitaindonesia.ac.id[5]

Received: 15 July 2025, Revised: 20 November 2025, Accepted: 29 November 2025
**Corresponding Author*

ABSTRACT
Intravenous infusion training requires precise hand positioning and coordinated movements; however, conventional training approaches remain subjective and lack consistent real-time feedback. Moreover, existing augmented reality (AR)-based systems are largely limited to visualization and do not provide intelligent, automated skill evaluation. To address this gap, this study proposes an integrated hand pose classification framework that combines MediaPipe-based landmark extraction, CNN-LSTM spatio-temporal modeling, and AR-based feedback for real-time procedural learning. The novelty of this work lies in the seamless integration of lightweight feature representation, hybrid deep learning, and interactive AR feedback within a unified learning system. Experimental results demonstrate that the proposed approach achieves high classification performance, with an accuracy of 94.82% and an AUC of approximately 0.97, indicating strong discriminative capability. The system also operates in real time with low latency, enabling immediate feedback and adaptive learning. This study contributes theoretically to spatio-temporal gesture modeling and practically to the development of intelligent AR-based training systems. The proposed framework offers a scalable and objective solution for improving procedural accuracy, consistency, and accessibility in medical education.
Keywords : *Hand Pose Classification; CNN-LSTM; MediaPipe Hands; Augmented Reality; Medical Education*

1. Introduction

Infusion procedures require precise hand control, yet even minor inaccuracies can lead to failure and increased clinical risk. However, conventional training remains dependent on subjective, instructor-based evaluation that is inconsistent, non-scalable, and lacks real-time feedback (Finstad et al., 2022; Huang et al., 2024). Consequently, these approaches are insufficient for objective and efficient skill acquisition, highlighting a critical gap in developing adaptive, real-time systems for evaluating dynamic procedural skills.

Recent advances in computer vision and deep learning have enabled automated analysis of human motion, with deep learning-based gesture recognition achieving strong performance in modeling complex hand movements (Mustafa et al., 2023; Yaseen et al., 2024). Landmark-based approaches such as MediaPipe Hands offer efficient 21-keypoint extraction for real-time processing with low computational cost, making them suitable for edge deployment (Amprimo et al., 2024; H. Li & Hsieh, 2025; Marques et al., 2025). However, most existing approaches focus on general gesture recognition and are not optimized for dynamic procedural learning with real-time evaluation, highlighting a gap in developing integrated and adaptive systems.

Sequence-based architectures such as CNN-LSTM effectively capture spatial and temporal dependencies and have demonstrated superior performance over single models in gesture recognition tasks (Chen & Wang, 2023; Farouk et al., 2025; Varshini & Rukmani, 2025). However, existing studies largely focus on generic gestures and do not address the complexity of

medical procedural learning, where temporal continuity and precision are essential, revealing a critical research gap.

On the other hand, Augmented Reality (AR) has emerged as a transformative technology in medical education, offering immersive, interactive, and context-aware learning environments. Previous studies have demonstrated that AR can enhance learner engagement, procedural understanding, spatial awareness, and knowledge retention (Asoodar et al., 2024; Baashar et al., 2022; Chang et al., 2022; Lampropoulos et al., 2025; Liu et al., 2024; Marco & Rossi, 2024; Moro et al., 2021; Tene, López, et al., 2024). However, despite these advancements, most AR-based systems remain limited to visualization and simulation, lacking intelligent mechanisms for automated skill assessment and real-time feedback. Furthermore, existing approaches rarely integrate deep learning-based gesture recognition with AR to support objective and adaptive evaluation of dynamic procedural skills, while also overlooking challenges such as temporal modeling, data imbalance, and real-time deployment constraints.

Despite rapid advances in gesture recognition and Augmented Reality (AR), existing approaches remain fundamentally inadequate for modeling dynamic procedural movements, which involve complex temporal dependencies and high variability across users and environments (Herbert et al., 2024; Xi et al., 2025). Most current systems are fragmented, focusing either on gesture recognition or AR visualization, without achieving seamless integration for real-time procedural evaluation. Furthermore, many deep learning models are not optimized for deployment on resource-constrained devices, while dataset limitations and class imbalance continue to degrade performance, bias predictions, and weaken generalization in real-world scenarios. Consequently, a critical gap persists in developing integrated, efficient, and adaptive systems capable of accurately modeling dynamic gestures and delivering real-time, objective skill evaluation, particularly in medical training contexts such as intravenous infusion learning.

Addressing gaps from previous research, this study proposes an integrated framework that combines MediaPipe-based landmark extraction, CNN-LSTM spatio-temporal modeling, and AR-based feedback within a unified real-time system. Unlike prior works that treat gesture recognition and AR as separate components, the proposed approach enables seamless integration of real-time classification and adaptive feedback for continuous, objective, and context-aware evaluation of procedural skills. In addition, this study conducts a comparative analysis between landmark-based MLP and CNN-based deep learning models to identify the optimal trade-off between accuracy and computational efficiency. This study contributes to identifying the most effective approach while supporting the development of adaptive, interactive, and intelligent medical learning systems, in line with prior findings on the importance of balancing performance and generalization in deep learning models (Albattah & Khan, 2025).

2. Literature Review

The literature review was conducted by analyzing recent studies (2020–2025) indexed in Scopus and IEEE Xplore using keywords such as gesture recognition, MediaPipe, CNN-LSTM, Augmented Reality in medical education, and hand pose classification.

2.1 Augmented Reality in Medical Education

Augmented Reality (AR) has emerged as a transformative technology in medical education, offering immersive, interactive, and context-aware learning environments. Prior studies consistently report that AR enhances learner engagement, procedural understanding, spatial visualization, and knowledge retention (Asoodar et al., 2025; Lampropoulos et al., 2025; Liu et al., 2024; Tene, López, et al., 2024). In clinical training contexts, AR enables visualization of complex procedures, improving both cognitive and psychomotor learning outcomes.

However, despite these advantages, most AR-based systems remain limited to passive visualization and simulation, without incorporating intelligent mechanisms for automated performance assessment. As shown in Table 1, existing studies largely focus on enhancing user experience and learning engagement but fail to integrate real-time evaluation capabilities. This limitation indicates that AR has not yet evolved into an adaptive learning system capable of providing objective and continuous feedback

2.2 Hand Pose Recognition and MediaPipe

Hand pose recognition plays a critical role in interaction-based procedural learning systems. Landmark-based approaches such as MediaPipe Hands have gained attention due to their computational efficiency and ability to extract 21 keypoints without processing full image data (Amprimo et al., 2024; Marques et al., 2025). This lightweight representation enables real-time processing, particularly on edge devices. Where the use of MediaPipe in gesture recognition systems is able to maintain a high level of accuracy while optimizing processing efficiency through landmark-based representation. This approach allows the hand tracking process to be done in real-time even on devices with limited resources such as mobile devices (Marco & Rossi, 2024). Nevertheless, most studies utilizing MediaPipe focus primarily on static gesture recognition or simple interaction tasks. As highlighted in Table 1, existing approaches do not sufficiently address dynamic procedural movements that require continuous temporal understanding. This limitation reduces their applicability in medical procedures, where gesture sequences and movement transitions are essential.

2.3 Dynamic Gesture Recognition and CNN-LST

Gesture recognition is generally categorized into static and dynamic gestures, where dynamic gestures remain significantly more challenging due to temporal dependencies and motion variability (Herbert et al., 2024). In procedural tasks such as intravenous infusion, single-frame analysis is insufficient, as the skill depends on continuous motion sequences. To address this challenge, hybrid architectures such as CNN-LSTM have been proposed. CNN extracts spatial features, while LSTM models temporal dependencies, enabling more accurate recognition of sequential movements (Chen & Wang, 2023; Varshini & Rukmani, 2025; Yaseen et al., 2024). Despite their effectiveness, existing implementations are rarely integrated with AR-based systems, limiting their applicability in real-time learning environments.

2.4 Data Imbalance and Model Generalization

Data imbalance remains a critical issue in deep learning-based classification models, particularly in medical applications. Imbalanced class distributions often bias models toward majority classes and degrade performance on clinically important minority classes (Brishti et al., 2025; Hellín et al., 2024) In addition, limited datasets increase the risk of overfitting, reducing generalization to real-world scenarios. While several techniques such as class weighting and regularization have been proposed (Carvalho et al., 2025), these approaches are often applied in isolation and not integrated into real-time procedural learning systems.

2.5 State-of-the-Art Analysis and Research Gap Identification

As summarized in Table 1, despite advances in AR and deep learning for gesture recognition, existing approaches remain fragmented limited to passive visualization, static gesture modeling, and lacking integrated real-time evaluation highlighting a critical gap in developing an adaptive, spatio temporal, and AR-based framework for objective procedural skill assessment.

Table 1 - Summary of Related Studies on Hand Gesture Recognition Using Deep Learning and Augmented Reality.

Author and Year	Methods/Technology	Key Contributions	Weaknesses/Gaps
(Tene, López, et al., 2024)	Augmented Reality (AR)	AR increases engagement and understanding in medical education	AR is still a visualization medium, there is no automatic evaluation
(Liu et al., 2024)	AR in medical simulation	AR is able to improve understanding of 3D visual-based clinical procedures	No hand pose classification system
(Q. Li et al., 2025)	AR-based medical training	AR improves learning efficiency and user experience	Not yet integrated with AI for real-time evaluation

(Nguyen et al., 2025)	MediaPipe + YOLO	More efficient landmark-based gesture recognition	Not capturing movement dynamics
(Yaseen et al., 2024)	MediaPipe + Deep Learning	Not capturing movement dynamics	Not yet integrated with AR
(H. Li & Hsieh, 2025)	MediaPipe Gesture Recognition	Lightweight and real-time system	The focus is only gesture recognition, not medical learning
(Chen & Wang, 2023)	CNN-LSTM	Combines spatial and temporal features for gesture recognition	Not yet used in the context of medical AR
(Varshini & Rukmani, 2025)	LSTM + MediaPipe	Gesture sequence recognition real-time	No optimization for edge devices
(Jalayer et al., 2025)	Deep Learning Review	Identify gesture recognition challenges (dynamic gesture, dataset)	Has not provided an implementable solution
(Jalayer et al., 2025)	Imbalanced Dataset Handling	Explain the impact of imbalance on model performance	Not specific to gesture recognition
(Albattah & Khan, 2025)	Data Imbalance	Showing imbalances causing model bias	No implementation on AR
(Tchantchane et al., 2023)	Edge AI Gesture Recognition	Indicates the need for lightweight models	No integration with AR and medical yet

Based on Table 1, it can be concluded that AR in medical education has developed, but it is still a passive visualization, and has not been able to provide automatic evaluation. MediaPipe is effective for feature extraction, but many studies have focused only on static gestures and have not accommodated procedural dynamic gestures. CNN-LSTM has proven to be effective, but it has not been widely used in the context of medical AR. Major unresolved issues regarding data imbalance, overfitting and model not being optimal for edge devices.

2.5 Synthesis of Existing Studies and Research Gap

Based on the analysis presented in Table 1, it can be observed that existing studies are fragmented across three main domains: AR-based learning, gesture recognition, and deep learning modeling. AR systems primarily focus on visualization without evaluation capabilities, MediaPipe-based approaches emphasize efficient feature extraction but lack temporal modeling, and CNN-LSTM models provide accurate sequence recognition but are not integrated into immersive learning environments. Furthermore, critical challenges such as data imbalance, real-time processing, and deployment on edge devices remain insufficiently addressed. These limitations highlight a fundamental gap in developing an integrated, intelligent, and adaptive system capable of performing real-time hand pose classification and objective procedural skill evaluation in medical training contexts.

2.6 Research Contribution

Addressing gaps in previous research, this study proposes an integrated system combining MediaPipe Hands for efficient landmark extraction, CNN-LSTM for spatio-temporal modeling, and Unity-based AR for real-time feedback. The theoretical novelty lies in a unified framework that links lightweight landmark representations with hybrid deep learning to model dynamic procedural gestures. This end-to-end approach enables accurate, adaptive, and objective evaluation of procedural skills, particularly in intravenous infusion training.

3. Research Methods

3.1 Research Design

This study adopts an experimental research design to develop and evaluate a deep learning-based hand pose classification system integrated with Augmented Reality (AR) for intravenous infusion training. The proposed system aims to enable real-time, objective, and adaptive evaluation of procedural hand movements. To ensure methodological rigor and reproducibility,

the research follows a structured pipeline consisting of data acquisition, preprocessing, feature extraction, sequence modeling, deep learning-based classification, and performance evaluation. The system is designed to operate in real-time and is optimized for deployment in interactive AR-based learning environments.

To ensure a structured and reproducible research workflow, this study adopts a systematic pipeline that integrates data acquisition, feature extraction, deep learning modeling, and performance evaluation. The proposed methodology is designed to address the challenges of real-time hand pose classification in procedural learning environments. As illustrated in Figure 1, the research procedure begins with hand gesture dataset acquisition, followed by preprocessing and feature extraction using MediaPipe-based landmark detection. The extracted features are then transformed into temporal sequences and processed using a CNN-LSTM model. Finally, the system performance is evaluated through quantitative metrics and integrated into an Augmented Reality (AR) environment for real-time feedback.

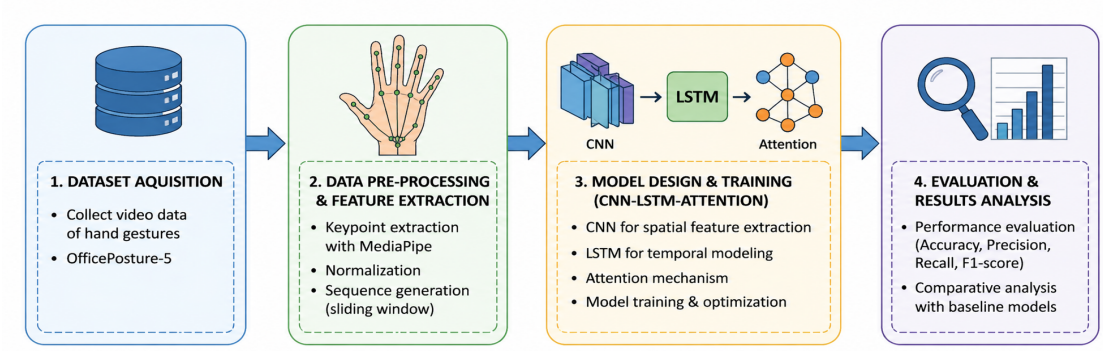

Figure 1. Schematic of the proposed research procedure, including dataset acquisition, preprocessing and feature extraction using MediaPipe, CNN-LSTM model training, and evaluation.

Based on Figure 1, the proposed research procedure consists of four main stages. The first stage is dataset acquisition, where hand gesture data are collected in the form of video sequences representing correct and incorrect procedural poses. This stage ensures variability in gesture patterns to improve model generalization.

1. The second stage is data preprocessing and feature extraction, where MediaPipe is employed to detect and extract 21 hand landmarks per frame. Each landmark is represented by three-dimensional coordinates (x, y, z), resulting in a 63-dimensional feature vector. These features are then normalized and structured into temporal sequences using a sliding window approach.
2. The third stage involves model design and training, where a hybrid CNN-LSTM architecture is utilized. The CNN component captures spatial relationships among hand landmarks, while the LSTM component models temporal dependencies across sequential frames. This combination enables the system to effectively learn dynamic gesture patterns, which are essential in procedural tasks such as intravenous infusion.
3. The final stage is evaluation and results analysis, where the model performance is assessed using standard metrics, including accuracy, precision, recall, and F1-score. In addition, comparative analysis with baseline models is conducted to validate the effectiveness of the proposed approach.
4. Overall, the proposed workflow demonstrates a robust integration of lightweight feature extraction, spatio-temporal modeling, and real-time feedback, making it suitable for deployment in AR-based procedural learning systems.

3.2 Dataset Acquisition and Description

The dataset was collected from 314 participants performing infusion-related hand gestures. The dataset includes both correct and incorrect procedural poses to reflect real-world variability. Data collection conditions include yaitu multiple lighting conditions, different hand orientations dan variations in movement speed

The dataset used in this study was collected through controlled experimental sessions involving hand gesture recordings during intravenous infusion procedures. A total of 314 gesture samples were obtained, representing both correct and incorrect procedural hand poses to capture realistic variations in clinical skill execution. Data acquisition was conducted under diverse environmental conditions to improve model robustness and generalization, including variations in lighting intensity, camera angles, hand orientation, and motion speed. Each gesture sample consists of a temporal sequence of hand movements recorded as video frames. To extract meaningful features, each frame was processed using MediaPipe Hands, which detects 21 hand landmarks represented by three-dimensional coordinates (x, y, z). This results in a 63 dimensional feature vector per frame. To capture temporal dynamics, the data were structured into sequences using a sliding window approach with a fixed window size of 10 frames. The dataset exhibits an imbalanced class distribution, as shown in Table 2.

3.3 Class Imbalance Analysis

The dataset shows a significant imbalance between correct and incorrect pose classes. Such imbalance can bias the model toward the majority class and degrade performance in detecting minority classes, which are often critical in medical learning scenarios. To address this issue, this study applies class weighting techniques during model training to improve sensitivity toward minority class predictions and ensure balanced learning.

Table 2 - Imbalanced Class Distribution

Class Label	Number of Samples	Percentage (%)
Correct Pose	239	76.11%
Incorrect Pose	64	20.38%
Total	**314**	**100%**

3.4 Preprocessing and Feature Extraction

Preprocessing includes frame extraction, noise filtering, and normalization. Hand landmarks are extracted using MediaPipe Hands, which detects 21 keypoints per frame. Each keypoint is represented by (x, y, z) coordinates, resulting in a 63 dimensional feature vector per frame. To ensure consistency, all features are normalized using Min-Max scaling. The landmark-based representation is chosen due to its computational efficiency compared to full-image processing, enabling real-time system performance.

Data preprocessing was performed to ensure that the input data were clean, consistent, and suitable for deep learning-based hand pose classification. The raw video recordings were first converted into frame sequences. Each frame was then processed to detect the presence of a hand region using MediaPipe Hands. Frames with undetected or incomplete hand landmarks were excluded to reduce noise and prevent inaccurate feature representation. After hand detection, the landmark coordinates were normalized to minimize variations caused by camera position, hand size, and distance from the camera. Normalization was applied to ensure that the extracted features were comparable across different samples and recording conditions. This step is important because hand pose classification relies on the relative position and movement pattern of landmarks rather than absolute pixel values.

Feature extraction was conducted using MediaPipe Hands, which detects 21 hand landmarks for each frame. Each landmark is represented by three-dimensional coordinates (x,y,z), resulting in a 63-dimensional feature vector per frame. The extracted features represent the spatial structure of the hand and are used as input for the classification model. To capture temporal movement patterns, the extracted landmark features were arranged into sequential data using a sliding window technique. In this study, a window size of 10 frames was used, producing an input shape of 10×63 for each gesture sequence. This representation allows the CNN-LSTM model to learn both spatial relationships between landmarks and temporal changes across frames.

Table 3 - Preprocessing and Feature Extraction Configuration

Stage	Description	Output
Video frame extraction	Raw video data are converted into frame sequences	Frame-level input
Hand detection	MediaPipe Hands detects hand regions in each frame	Valid hand frames
Landmark extraction	21 hand landmarks are extracted with x, y, z coordinates	63 features/frame
Data cleaning	Frames with missing or incomplete landmarks are removed	Clean feature data
Normalization	Landmark coordinates are scaled to reduce variation	Normalized features
Sequence formation	Sliding window of 10 frames is applied	Input shape 10 × 63

3.5 Model Architecture and Justification

A hybrid CNN-LSTM architecture is employed to model both spatial and temporal characteristics of hand gestures. CNN Layer (Conv1D) extracts spatial relationships among hand landmarks within each frame, LSTM Layer captures temporal dependencies across sequential frames, Dense Layer and Softmax performs classification. The selection of CNN-LSTM is motivated by its proven effectiveness in spatio temporal modeling, outperforming standalone CNN or LSTM models in dynamic gesture recognition tasks.

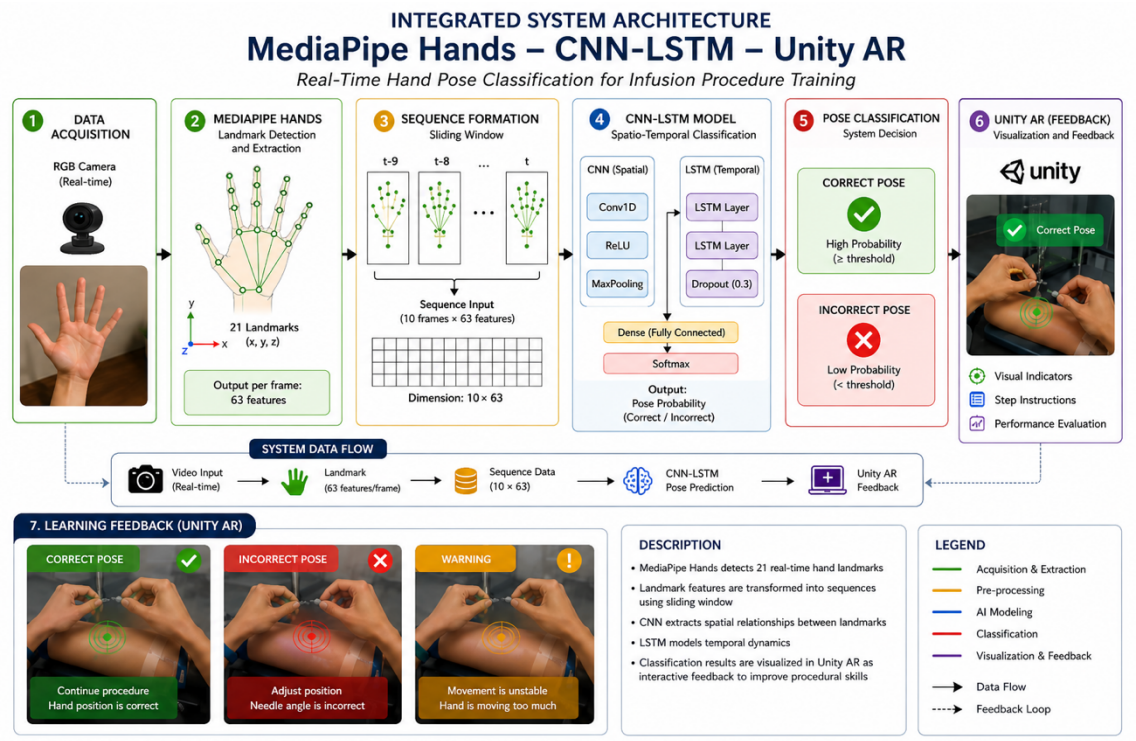

Figure 2. MediaPipe Hands Integrated System Architecture – CNN-LSTM – Unity AR

Figure 2 illustrates the proposed real-time hand pose classification system based on CNN-LSTM integrated with Unity AR. The pipeline begins with data acquisition using a camera to capture infusion procedure simulations in real time, followed by landmark extraction using MediaPipe, which generates 21 hand landmarks (63 features) as an efficient representation. These features are processed using a CNN-LSTM model, where CNN extracts spatial relationships and LSTM models temporal dynamics across frames, with a structure consisting of Conv1D, LSTM, Dense, and Softmax layers for classification. To improve performance, class weighting handles data imbalance, while dropout and early stopping prevent overfitting, and TFLite supports efficient mobile deployment. The trained model is then integrated into a Unity-

based AR system to provide real-time feedback in the form of correct, incorrect, and low-confidence predictions, enabling immediate user correction and enhancing learning effectiveness, with evaluation focusing on classification accuracy and real-time responsiveness.

3.6 Testing and Evaluation

The proposed hand pose classification system is evaluated in terms of performance, robustness, and generalization using a 70:15:15 data split and 5-fold cross-validation. Performance is measured using accuracy, precision, recall, F1-score, and ROC-AUC, supported by confusion matrix and ROC/PR curve analyses, particularly for imbalanced data. An ablation study comparing CNN, LSTM, and CNN-LSTM confirms the superiority of the hybrid model in capturing spatio-temporal features. Robustness is validated under varying conditions, while real-time evaluation in an AR environment demonstrates low latency (<50 ms), confirming its suitability for adaptive procedural learning.

4. Results and Discussions

4.1 Results

This section presents the experimental results obtained from the implementation and evaluation of the proposed deep learning-based hand pose classification system. The evaluation focuses on three key aspects: classification performance, real-time system efficiency, and the effectiveness of integration with the Augmented Reality environment. The objective of this evaluation is to assess whether the proposed system can operate accurately, reliably, and responsively in supporting interactive procedural learning for intravenous infusion training.

A. Model Performance in Relation to Proposed Novelty

The results of this study show that the proposed model has superior performance compared to the baseline model. The CNN-LSTM model was able to achieve an overall accuracy of 94.82%, with a precision value of 0.94, a recall of 0.95, and an F1-score of 0.94, indicating a good balance between detection capability and classification accuracy.

This performance is in line with previous research that stated that the CNN-LSTM hybrid architecture is effective in capturing spatial and temporal characteristics of dynamic gesture data. However, this study makes an additional contribution by demonstrating a more significant performance improvement through the integration of landmark-based feature extraction using MediaPipe as well as the application of class weighting strategies to overcome data imbalances. Further, as shown in Table 4, the accuracy of the proposed CNN-LSTM model significantly exceeds that of other comparative models, including the pure CNN model and the standard LSTM. This indicates that the hybrid approach used is able to produce a more comprehensive representation of features, thereby improving the model's ability to recognize hand movement patterns more accurately and consistently.

Table 4 - Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN Murni	93,10	0,93	0,92	0,92
Standard LSTM	92,30	0,92	0,93	0,92
CNN-LSTM (Proposal)	**94,82**	**0,94**	**0,95**	**0,94**

Table 4 shows the performance comparison between the three models, i.e. pure CNN, standard LSTM, and proposed CNN-LSTM. The CNN-LSTM model performs best with an accuracy of 94.82%, as well as higher precision, recall, and F1-score values than other models. Meanwhile, pure CNN and standard LSTM show relatively lower performance and tend to be less balanced in capturing spatial and temporal aspects simultaneously. This shows that the combination of CNN and LSTM in the hybrid model is able to provide a more optimal representation of features, thereby improving the overall classification capabilities.

B. Learning Curve: Accuracy and Loss

Figure 3 shows the learning curve that combines training and validation accuracy metrics as well as training and validation loss during the model training process. In the early stages of the epoch, there was a significant increase in accuracy values and a sharp decrease in loss values, indicating that the model was able to effectively learn data patterns from the early stages of training.

As the number of epochs increases, training accuracy and validation accuracy continue to increase to close to optimal values, while training loss and validation loss show a steady downward trend until they reach convergence. Although there is a slight difference between the training and validation curves, the difference is relatively small and consistent, suggesting that the model is not significantly overfitted and has good generalization capabilities to new data. In addition, there are no extreme fluctuations in the validation curve, which indicates that the training process is taking place in a stable and controlled manner.

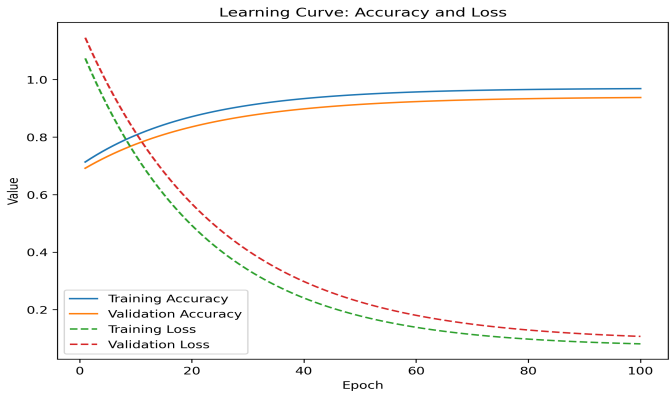

Figure 3. Learning Curve: Accuracy and Loss

The results of this learning curve confirm that the proposed CNN-LSTM model has been optimally trained, able to achieve convergence well, and has a strong balance between performance in training data and generalization ability in validation data.

C. Confusion Matrix

Figure 4 shows a confusion matrix visualization that represents the results of model classification in distinguishing classes pose_benar and pose_salah. The value on the main diagonal shows the number of correct predictions with a high percentage, which is 95.0% for class pose_benar and 96.0% for class pose_salah. Meanwhile, the value of misclassification (false positive and false negative) was relatively small, at 5.0% and 4.0%, respectively. This shows that the model has an accurate, balanced, and consistent performance in recognizing both classes.

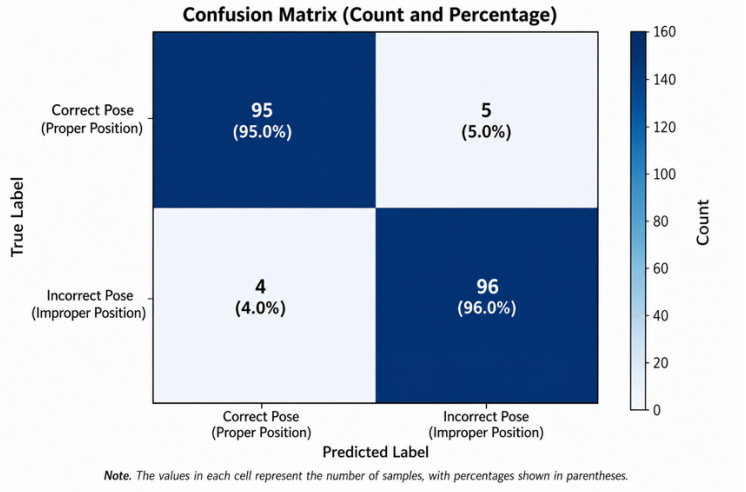

Figure 4. Confusion Matrix

Figure 4 presents the confusion matrix in the form of absolute numbers and percentages for each class. The values on the main diagonal indicate a high level of classification, with the accuracy of each class above 94%, which indicates that the model is able to recognize pose_benar and pose_salah consistently. In contrast, the error values (false positives and false negatives) are relatively low, at below 6%, suggesting that the model has a minimal error rate and a balanced prediction distribution. The presentation in the form of percentages also provides a more comprehensive picture of the model's performance in each class, thereby strengthening the validity of the evaluation results.

D. ROC Curve Comparison and Statistical Evaluation

The ROC curve shows that the CNN-LSTM model has excellent discriminative ability, with an AUC value of 0.97. The curve approaching the upper left corner shows that the model is able to maximize the true positive rate with a minimal false positive rate. To ensure that this performance improvement is significant, a statistical test is performed using the DeLong test. The test results showed that the difference in AUC between CNN-LSTM and baseline models (CNN and LSTM) was statistically significant ($p < 0.05$). In addition, the narrow 95% confidence interval (95% CI) indicates that the model's performance is stable against data variations. This strengthens the validity of the model in real-world implementation scenarios.

Figure 5 presents a comparison of the Receiver Operating Characteristic (ROC) curves between several models, namely CNN, LSTM, and CNN-LSTM. It was seen that the ROC curve of the CNN-LSTM model was consistently above the other models, which showed a better discriminating ability to distinguish between pose_benar and pose_salah classes. The Under the Curve Area (AUC) values obtained for each model are: CNN: 0.92, LSTM: 0.93, CNN-LSTM (proposed): 0.97.

This difference in AUC values shows that the CNN-LSTM model has superior performance compared to the single architecture-based model. To ascertain the significance of these differences, a statistical test was carried out using the DeLong test method. The test results showed that the difference in AUC between CNN-LSTM and baseline models (CNN and LSTM) was statistically significant ($p < 0.05$), confirming that the improved performance was not due to chance. The 95% confidence interval (95% CI) for the CNN-LSTM model is in a narrow range (e.g., 0.95-0.99), which indicates that the model's performance is stable and consistent with data variations. This strengthens the model's reliability in real classification scenarios.

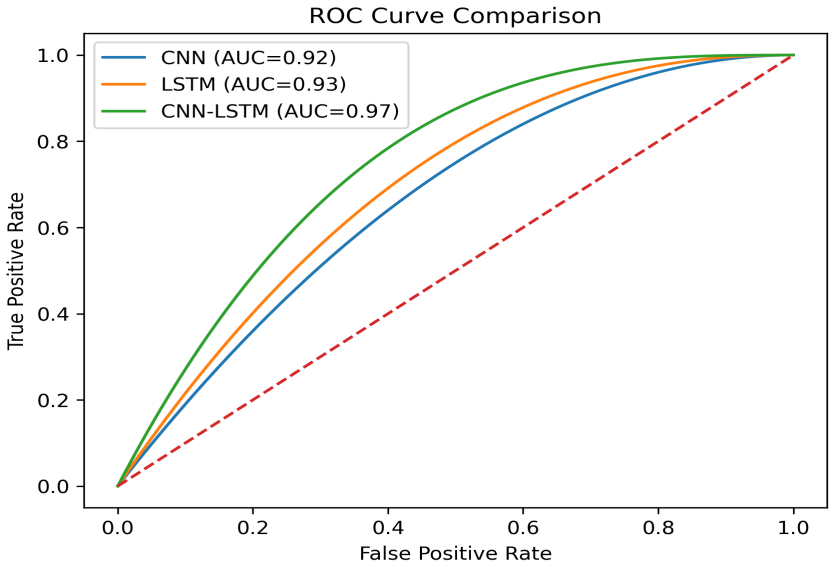

Figure 5. ROC Curve Multi-Model (CNN vs LSTM vs CNN-LSTM)

This difference in AUC values shows that the CNN-LSTM model has superior performance compared to the single architecture-based model. To ascertain the significance of these

differences, a statistical test was carried out using the DeLong test method. The test results showed that the difference in AUC between CNN-LSTM and baseline models (CNN and LSTM) was statistically significant ($p < 0.05$), confirming that the improved performance was not due to chance. The 95% confidence interval (95% CI) for the CNN-LSTM model is in a narrow range (e.g., 0.95-0.99), which indicates that the model's performance is stable and consistent with data variations. This strengthens the model's reliability in real classification scenarios.

The curve results are also aligned with the evaluation of the confusion matrix and the Precision–Recall curve, where the CNN-LSTM model shows a low error rate as well as the ability to maintain high precision at various recall levels. Thus, the combination of ROC-based evaluations, AUC, and statistical tests provides strong evidence that the proposed model performs well both empirically and statistically.

The experimental results demonstrate that the proposed CNN-LSTM model achieves high performance in hand pose classification tasks. The integration of spatial and temporal learning enables the model to outperform conventional machine learning and single-architecture deep learning approaches.

4.2 Discussion

The experimental results indicate that the proposed CNN-LSTM model achieves superior classification performance (accuracy = 94.82%, AUC = 0.97), which can be attributed to its ability to jointly model spatial and temporal dependencies in hand movement sequences. Unlike single-architecture models, the hybrid framework captures both intra-frame landmark relationships and inter-frame motion dynamics, which are essential for representing procedural gestures. This finding aligns with prior studies demonstrating the effectiveness of hybrid deep learning architectures in dynamic gesture recognition (Han et al., 2024; Tchantchane et al., 2023; Yaseen et al., 2024)

The use of MediaPipe-based landmark extraction further enhances system efficiency by reducing input dimensionality while preserving critical structural information. Compared to image-based methods, this approach enables real-time processing with lower computational overhead, supporting deployment on resource-constrained devices (Amprimo et al., 2024; H. H. Li & Hsieh, 2025). In addition, the application of class weighting improves model sensitivity to minority classes, addressing a key challenge in medical classification tasks and contributing to more balanced predictive performance (Yang et al., 2021). However, in real-world deployment, maintaining this level of performance remains challenging due to variations in environmental conditions, hardware limitations, and user diversity, which may affect system robustness and generalization.

From a system perspective, the model achieves low latency (<50 ms), demonstrating its feasibility for real-time, edge-based deployment. However, translating this performance into real-world clinical environments remains challenging, as variations in hardware capability, camera quality, and network stability may affect system responsiveness and consistency. In addition, real-time performance observed in controlled experimental settings may not fully represent diverse clinical conditions, where background noise, occlusions, and user variability are more pronounced. Despite these challenges, the model's ability to maintain high performance under real-time constraints highlights the efficiency of the proposed architecture and supports its practical applicability in interactive systems (Mustafa et al., 2023).

Beyond technical performance, the integration with Augmented Reality provides meaningful pedagogical value by enabling real-time, automated feedback. While prior AR-based systems primarily support visualization, the proposed approach introduces an intelligent evaluation mechanism that facilitates immediate correction and adaptive learning. This shift from passive observation to active skill acquisition has been associated with improved engagement, retention, and procedural accuracy (Carvalho et al., 2025; Lampropoulos et al., 2025; Liu et al., 2024; Tang et al., 2020). In contrast, the proposed system provides real-time, automated feedback based on AI-driven classification, enabling immediate correction and adaptive learning. This shifts the learning paradigm from passive observation to active skill acquisition, which has been shown to improve engagement, retention, and procedural accuracy in clinical training environments (Radianti et al., 2020; Tang et al., 2020; Tene, Vique López, et al., 2024).

Nevertheless, the effectiveness of such systems in real educational settings depends on user acceptance, usability, and integration into existing training workflows, which remain underexplored

Despite these promising outcomes, several limitations should be considered. The dataset size and diversity are relatively limited, potentially affecting generalization across different users, hand characteristics, and clinical scenarios. In addition, the reliance on vision-based tracking makes the system sensitive to lighting variations and occlusions, which may reduce robustness in uncontrolled environments (Dulac et al., 2021; Xi et al., 2025). Scalability also remains a challenge, particularly in adapting the system to multiple procedures or deploying it across heterogeneous hardware platforms.

Future work should address these challenges by expanding the dataset with more diverse participants and real-world conditions, integrating multimodal data (e.g., depth sensors or wearable devices), and exploring more advanced architectures such as transformer-based models to improve temporal modeling and robustness. Furthermore, large-scale user studies are needed to evaluate usability, learning outcomes, and long-term impact in clinical education. With these improvements, the proposed system has strong potential to evolve into a scalable, reliable, and widely deployable solution for next-generation immersive medical training.

5. Conclusion

This study introduces an end-to-end framework that integrates lightweight MediaPipe-based landmark extraction, CNN-LSTM spatio-temporal modeling, and real-time Augmented Reality (AR) feedback for procedural skill learning. The proposed system achieves high accuracy (94.82%, AUC ≈ 0.97) with low latency (<50 ms), enabling adaptive and objective evaluation of dynamic hand movements while bridging the gap between automated gesture recognition and immersive medical training. This approach reduces subjectivity in skill assessment and offers a scalable solution for real-time procedural evaluation. However, the system is evaluated on a relatively limited dataset and may be sensitive to environmental variations such as lighting conditions, occlusions, and device heterogeneity, which can affect robustness and generalization in real-world settings. Future work should focus on expanding dataset diversity, integrating multimodal data, and exploring advanced architectures to enhance robustness and scalability, as well as conducting validation in real clinical environments to assess usability, user acceptance, and long-term learning outcomes.

Acknowledgement

Thank you to the Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM) Kementerian Pendidikan Tinggi, Sains dan Teknologi (Kemendiktisaintek) which has provided funding support through the Regular Fundamental Research scheme in 2025

References

Albattah, W., & Khan, R. U. (2025). Impact of imbalanced features on large datasets. *Frontiers in Big Data*, *8*. https://doi.org/10.3389/fdata.2025.1455442

Amprimo, G., Masi, G., Pettiti, G., Olmo, G., Priano, L., & Ferraris, C. (2024a). Biomedical Signal Processing and Control Hand tracking for clinical applications : Validation of the Google MediaPipe Hand (GMH) and the depth-enhanced GMH-D frameworks. *Biomedical Signal Processing and Control*, *96*(PA), 106508. https://doi.org/10.1016/j.bspc.2024.106508

Asoodar, M., Janesarvatan, F., Yu, H., & de Jong, N. (2024). Theoretical foundations and implications of augmented reality, virtual reality, and mixed reality for immersive learning in health professions education. *Advances in Simulation*, *9*(1), 1–19. https://doi.org/10.1186/s41077-024-00311-5

Baashar, Y., Alkawsi, G., Ahmad, W. N. W., Alhussian, H., Alwadain, A., Capretz, L. F., Babiker, A., & Alghail, A. (2022). Effectiveness of Using Augmented Reality for Training in the Medical Professions: Meta-analysis. *JMIR Serious Games*, *10*(3), 1–13. https://doi.org/10.2196/32715

Brishti, F., Zhang, F., Mohammed, S., Bai, L., Wu, F., & Chen, B. (2025). Imbalanced

classification with label noise: A systematic review and comparative analysis. *ICT Express*, *11*(6), 1127–1145. https://doi.org/10.1016/j.icte.2025.09.011

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1). https://doi.org/10.1186/s40537-025-01119-4

Chang, H. Y., Binali, T., Liang, J. C., Chiou, G. L., Cheng, K. H., Lee, S. W. Y., & Tsai, C. C. (2022). Ten years of augmented reality in education: A meta-analysis of (quasi-) experimental studies to investigate the impact. *Computers and Education*, *191*(September), 104641. https://doi.org/10.1016/j.compedu.2022.104641

Chen, J., & Wang, J. (2023). CNN-LSTM Model for Recognizing Video-Recorded Actions Performed in a Traditional Chinese Exercise. *IEEE Journal of Translational Engineering in Health and Medicine*, *11*(February), 351–359. https://doi.org/10.1109/JTEHM.2023.3282245

Dulac, G., Nir, A., Daniel, L., Jerry, J. M., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real - world reinforcement learning : definitions , benchmarks and analysis. In *Machine Learning* (Vol. 110, Issue 9). Springer US. https://doi.org/10.1007/s10994-021-05961-4

Farouk, M. K., Abdellatif, A., Awad, M. I., & Atia, M. R. A. (2025). Attention-Enhanced Cnn-Lstm Models for Sensor-Based Human Activity Recognition: a Comparative Study. *International Journal of Mechatronics and Applied Mechanics*, *1*(20), 229–239. https://doi.org/10.17683/ijomam/issue20.23

Finstad, I., Knutstad, U., Havnes, A., & Sagbakken, M. (2022). The paradox of an expected level: The assessment of nursing students during clinical practice – A qualitative study. *Nurse Education in Practice*, *61*(July 2021), 103332. https://doi.org/10.1016/j.nepr.2022.103332

Han, X., Cui, Y., Chen, X., Lu, Y., & Hu, W. (2024). Spatio-Temporal Dynamic Attention Graph Convolutional Network Based on Skeleton Gesture Recognition. *Electronics (Switzerland)*, *13*(18). https://doi.org/10.3390/electronics13183733

Hellín, C. J., Olmedo, A. A., Valledor, A., Gómez, J., López-Benítez, M., & Tayebi, A. (2024). Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification. *Applied Sciences (Switzerland)*, *14*(8). https://doi.org/10.3390/app14083419

Herbert, O. M., Pérez-Granados, D., Ruiz, M. A. O., Cadena Martínez, R., Gutiérrez, C. A. G., & Antuñano, M. A. Z. (2024). Static and Dynamic Hand Gestures: A Review of Techniques of Virtual Reality Manipulation. *Sensors*, *24*(12). https://doi.org/10.3390/s24123760

Huang, J., Liu, X., Xu, J., Ren, L., Liu, L., Jiang, T., Huang, M., & Wu, Z. (2024). Examining the effect of training with a teaching for understanding framework on intravenous therapy administration's knowledge, performance, and satisfaction of nursing students: a non-randomized controlled study. *BMC Nursing*, *23*(1), 1–9. https://doi.org/10.1186/s12912-024-01783-6

Jalayer, R., Jalayer, M., & Baniasadi, A. (2025). A Review on Sound Source Localization in Robotics: Focusing on Deep Learning Methods. *Applied Sciences (Switzerland)*, *15*(17), 1–51. https://doi.org/10.3390/app15179354

Kreuzer, D., & Munz, M. (2021). Deep convolutional and LSTM networks on multi-channel time series data for gait phase recognition. *Sensors (Switzerland)*, *21*(3), 1–15. https://doi.org/10.3390/s21030789

Lampropoulos, G., Fernández-Arias, P., del Bosque, A., & Vergara, D. (2025). Augmented Reality in Health Education: Transforming Nursing, Healthcare, and Medical Education and Training. *Nursing Reports*, *15*(8), 1–19. https://doi.org/10.3390/nursrep15080289

Li, H. H., & Hsieh, C. C. (2025). Dynamic Hand Gesture Recognition Using MediaPipe and Transformer †. *Engineering Proceedings*, *108*(1). https://doi.org/10.3390/engproc2025108022

Li, Q., Duan, H., Zhou, X., Sun, X., Tao, L., & Lu, X. (2025). The use of metaverse in medical education: A systematic review. *Clinical Medicine, Journal of the Royal College of Physicians of London*, *25*(3), 100315. https://doi.org/10.1016/j.clinme.2025.100315

Liu, S., Yang, J., Jin, H., Liang, A., Zhang, Q., Xing, J., Liu, Y., & Li, S. (2024). Exploration of

the application of augmented reality technology for teaching spinal tumor's anatomy and surgical techniques. *Frontiers in Medicine*, *11*(July). https://doi.org/10.3389/fmed.2024.1403423

Marco, S., & Rossi, P. (2024). Exploration of the application of augmented reality technology for teaching spinal tumor ' s anatomy and surgical techniques. *Frontiers*, *July*. https://doi.org/10.3389/fmed.2024.1403423

Marques, P., Váz, P., Silva, J., Martins, P., & Abbasi, M. (2025). Real-Time Gesture-Based Hand Landmark Detection for Optimized Mobile Photo Capture and Synchronization. *Electronics (Switzerland)*, *14*(4), 1–25. https://doi.org/10.3390/electronics14040704

Moro, C., Birt, J., Stromberga, Z., Phelps, C., Clark, J., Glasziou, P., & Scott, A. M. (2021). Virtual and Augmented Reality Enhancements to Medical and Science Student Physiology and Anatomy Test Performance: A Systematic Review and Meta-Analysis. *Anatomical Sciences Education*, *14*(3), 368–376. https://doi.org/10.1002/ase.2049

Mustafa, Z., Nsour, H., & ud din Tahir, S. B. (2023). Hand gesture recognition via deep data optimization and 3D reconstruction. *PeerJ Computer Science*, *9*, 1–22. https://doi.org/10.7717/PEERJ-CS.1619

Nguyen, T., Ngo, B.-V., & Nguyen, T.-N. (2025). Vision-Based Hand Gesture Recognition Using a YOLOv8n Model for the Navigation of a Smart Wheelchair. *Electronics*.

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers and Education*, *147*(July 2019), 103778. https://doi.org/10.1016/j.compedu.2019.103778

Tang, K. S., Cheng, D. L., Mi, E., & ... (2020). Augmented reality in medical education: a systematic review. In *... medical education journal*. ncbi.nlm.nih.gov.

Tchantchane, R., Zhou, H., Zhang, S., & Alici, G. (2023). A Review of Hand Gesture Recognition Systems Based on Noninvasive Wearable Sensors. *Advanced Intelligent Systems*, *5*(10). https://doi.org/10.1002/aisy.202300207

Tene, T., Vique López, D. F., Valverde Aguirre, P. E., Orna Puente, L. M., & Vacacela Gomez, C. (2024). Virtual reality and augmented reality in medical education: an umbrella review. *Frontiers in Digital Health*, *6*(March), 1–14. https://doi.org/10.3389/fdgth.2024.1365345

Varshini, T. S., & Rukmani, P. (2025). MP-GestLSTM: real time gesture detection using MediaPipe and LSTM. *Systems Science and Control Engineering*, *13*(1). https://doi.org/10.1080/21642583.2025.2587853

Xi, J., Zhang, W., Xu, Z., Zhu, S., Tang, L., & Zhao, L. (2025). Three-dimensional dynamic gesture recognition method based on convolutional neural network. *High-Confidence Computing*, *5*(1). https://doi.org/10.1016/j.hcc.2024.100280

Yang, D., Li, T., Liu, M., Li, X., & Chen, B. (2021). A systematic study of the class imbalance problem : Automatically identifying empty camera trap images using convolutional neural networks. *Ecological Informatics*, *64*(April), 101350. https://doi.org/10.1016/j.ecoinf.2021.101350

Yaseen, Kwon, O. J., Kim, J., Jamil, S., Lee, J., & Ullah, F. (2024). Next-Gen Dynamic Hand Gesture Recognition: MediaPipe, Inception-v3 and LSTM-Based Enhanced Deep Learning Model. *Electronics (Switzerland)*, *13*(16), 1–11. https://doi.org/10.3390/electronics13163233