Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Emotion-Aware AI and Digital Transformation of Visual Culture 2025 6(5s)

DEEP LEARNING FOR GESTURE ANALYSIS IN PERFORMANCE TRAINING

Dr. Anusha Sreeram [1], Pratibha Sharma [2], Kairavi Mankad [3], Kalpana Rawat [4], Rahul Kumar Sharma [5], Milind Patil [6]

[1] Faculty of Operations and IT, ICFAI Business School (IBS), The ICFAI Foundation for Higher Education (IFHE), (Deemed to be University u/s 3 of the UGC Act 1956), Hyderabad, India
[2] Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India
[3] Assistant Professor, Department of Fashion Design, Parul Institute of Design, Parul University, Vadodara, Gujarat, India
[4] Assistant Professor, School of Business Management, Noida International University, Inida
[5] Assistant Professor, Department of Computer Science and Engineering, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India
[6] Department of E and TC Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India

Received 13 May 2025
Accepted 17 August 2025
Published 28 December 2025

Corresponding Author
Dr. Anusha Sreeram,
seeramanusha@gmail.com

DOI
10.29121/shodhkosh.v6.i5s.2025.6904

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Gesture analysis is an important element in performance training in various areas including dance, sports, theatre, music and rehabilitation, whereby accurate movement of the body, time and expressiveness is important to the overall quality. The method of traditional gesture evaluation is biased on professional observation, and this method is subjective, cumbersome and cannot be scaled. This paper describes a deep learning-based gesture recognition system in the area of performing training to provide objective and data-driven feedback and individualized skills acquisition. The suggested method combines pose estimation with the help of computer vision and the deep neural framework, such as Convolutional Neural Networks (CNNs) to extract spatial features and Long Short-Term Memory (LSTM) or Transformer models to model temporal motion. Multi-dimensional gesture characteristics like joint paths, velocity signals, symmetry, balance and rhythmical consistency are trained directly by video sequences without any physical feature engineering. The framework facilitates real time and offline analysis that enables performers to get corrective feedback in real time or performance longitudinally. Empirical analyses show that deep learning models are far more efficient than the traditional machine learning methods on accuracy of gesture identification, detection of temporal alignment, movement quality evaluation. It is also possible to score expressiveness, coordination and consistency quantitatively with the help of the system, which helps to train and measure progress. The proposed deep learning-based gesture analysis framework should have significant potential in intelligent performance training systems, online coach environments, and simulated learning environments, owing to the subjectivity reduction and increased accessibility.

Keywords: Deep Learning, Gesture Analysis, Performance Training, Pose Estimation, Temporal Modeling, Motion Analytics, Intelligent Coaching

1. INTRODUCTION

Gesture training is an important part of performance training in areas like dance, sports, music, theatre, and physical rehabilitation, where performance is directly affected by the quality of movement, timing, coordination and

expressiveness. Gestures streamline intricate spatio-temporal information, which entails the articulation of the joints, body position, rhythm, and dynamic actions and are the main indicators of skill mastery and performance proficiency. Gesture assessment is traditionally performed using the methods of expert observation and manual feedback in the traditional training environment, when a coach, instructor, or the person is required to visually observe movements and offer corrective feedback. Although this strategy has the advantage of being domain-aware and contextual, it is subjective and labor-intensive in nature, and could not be easily applied to large or distributed learning environments Dewi et al. (2023), Kang et al. (2022). Besides, the minute differences in movement dynamics or sampling consistency can be ignored, which constrains the granularity and objectivity of performance evaluation Qiang et al. (2021). Traditional methods of gesture recognition commonly use rule-based techniques, hand-crafted techniques, or simple motion capture techniques, which are difficult to exploit the complexity of the human movement. These algorithms usually rely on some fixed kinematic rules or threshold measures or shallow machine learning models that need substantial feature engineering and domain adaptation Yaseen et al. (2025). They are therefore weak to the changes in the style of performers, body structure, the camera perspective and the surrounding environment. Furthermore, the traditional systems are not as flexible as such it is not easy to offer personalized feedback or generalize in different contexts on performance like expressive dance and high-speed athletic gestures Mujahid et al. (2021). This makes them not useful in real world training conditions because they are not able to model long-term temporal dependencies and subtle motion transitions Knights et al. (2021).

The latest developments in deep learning have changed the paradigm of movement and performance analytics by allowing the learning of complex gesture representations using data, either raw video or sensor data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) models, and Transformer-based networks have proven very successful in modeling both spatial arrangements, temporal relationships and contextual relationships that are inherent to human motion Zhang et al. (2025). Through joint estimation and deep temporal modeling, the methods are applicable in tracking the joint movements accurately, motion fragility, rhythm stability, and coordination scheme in the absence of design of features manually. Scalable, real-time gesture recognition is also assisted by deep learning, presents fresh opportunities to intelligent coaching systems, automated performance evaluation, and dynamic training systems Hax et al. (2024), Zhang et al. (2025). The overall aim of the study is to create a deep learning-oriented structure of gesture recognition during training in performance that would overcome the drawbacks of the conventional evaluation tools. The suggested solution is expected to provide objective, interpretable, and fine-grained evaluation of gestures quality and facilitate real-time feedback and performance monitoring in the long term. The contributions of this work are majorly the incorporation of pose spatial feature with sophisticated temporal modeling, development of quantitative gesture quality measures and systematic testing on various performance conditions. The study performs a contribution leading to more reachable, scalable, and information driven training ecosystems that complement human competence instead of substituting them by linking computational intelligence with embodied performance analysis Chang et al. (2023).

2. RELATED WORK

Gesture recognition through vision has been a widely researched fundamental issue in human-computer interaction, and motion analysis and has been approached early on by handcrafted visual features like optical flow, silhouette contours, motion history images, and trajectory-based features of RGB or depth videos. Those were generally classical classifiers, such as Hidden Markov Models (HMMs), Support Vector Machines (SVMs) and k-Nearest Neighbors in order to identify pre-defined gesture categories Hax et al. (2024). Although these methods proved to be practical in controlled settings, they started degrading very quickly with the changes in lighting, background clutter, viewpoint variations, occlusions, and diversity in performers. Furthermore, the rule-based and feature-based pipelines involved high domain knowledge and were not flexible in moving to new performance environments Zhang et al. (2025).

The development of deep learning has radically changed pose and motion analysis with learning hierarchical features directly on visual information. Convolutional Neural Networks (CNNs) have become popular in the areas of spatial feature detecting and pose prediction using 2D or 3D video data, precisely positioning the body joints and skeletal systems based on monocular or multi-view videos. Recurrent neural network models (including: Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs)) have also been combined with CNN-based back-ends to model time-varying dependencies and motion change with time Chang et al. (2023). Transformer-based models and graph convolutional networks (GCNs) have also been viewed as a more recent origin of such models with the potential

to capture the long-range temporal dependencies and skeleton structure in more generally challenging motion tasks, resulting in increased robustness and generalization Miah et al. (2023). The analysis of gestures has also been used at various levels of performance training. In sports analytics, deep learning models can be applied to analyze the techniques of athletes, identify the presence of biomechanical inefficiencies, and prevent injuries with the help of movement patterns. The system of dance training uses pose-based models to evaluate the accuracy of choreography, the correctness of rhythms, and the quality of expression and provide automated feedback to the learner. Gesture analysis is used in music performance to aid in the interpretation of conductor movement, instrumental technique analysis, and expressive timing in music performance. Gesture and motion analytics are used to study rehabilitation, where they use this technology to keep track of motor recovery, evaluate compliance with therapy, and develop exercise routines tailored to the individual needs of patients with neurological or musculoskeletal issues Agab and Chelali (2023). The domain studies below show that deep learning-based gesture analysis can be generalized to many different tasks, but are often limited to specific tasks or highly controlled databases.

Even with this substantial development there are a number of gaps and challenges in research within present studies. Most techniques focus more on accuracy in gesture classification, and ignore qualitative aspects of gesture like expressiveness, fluidity and style variation, which are essential in performance training. The limitations in the datasets such as small number of samples, the dataset is not diverse, and performance quality metrics are not thoroughly annotated limit the generalizability of the model across populations and domains. On-the-fly deployment is still a difficult task because it may require a great deal of computation time, as well as latency and hardware needs, especially when using resource-sensitive training systems. Moreover, the majority of deep learning models are black boxes, which are not very interpretable, which makes users less trustful of them and does not allow them to get meaningful pedagogical feedback Chen et al. (2022). Ensuring explainability, personalization, ethical use of data, and cross-domain transfer learning are necessary to make further progress in gesture analysis based on recognition-focused systems to comprehensive, intelligent performance training systems Feng et al. (2025). This Table 1 brings to light tendencies, contributions, and ongoing issues in the current gesture analysis studies in performance training fields.

Table 1

Table 1 Summary of Related Work on Gesture Analysis in Performance Training

Ref.	Application Domain	Input Modality	Core Technique	Temporal Modeling	Key Contribution	Limitation
Hax et al. (2024)	General HCI	RGB Video	Handcrafted features + SVM	HMM	Early vision-based gesture recognition framework	Poor robustness to noise and viewpoint changes
Zhang et al. (2025)	Human Motion Analysis	RGB / Depth	Optical flow, silhouettes	Limited	Demonstrated feasibility of visual gesture modeling	High dependency on manual feature engineering
Chang et al. (2023)	Pose Estimation	RGB Video	CNN-based pose estimation	LSTM	Accurate joint localization with temporal consistency	Struggles with occlusion and fast motions
Miah et al. (2023)	Action Recognition	Skeleton Data	GCN + Transformer	Transformer	Captured long-range spatio-temporal dependencies	High computational complexity
Agab and Chelali (2023)	Sports Training	RGB / Skeleton	CNN–LSTM	LSTM	Technique evaluation and injury risk analysis	Domain-specific, limited generalization
Chen et al. (2022)	Dance Performance	RGB Video	Pose-based deep models	LSTM	Automated choreography and rhythm assessment	Expressiveness modeling remains weak
Feng et al. (2025)	Music Performance	RGB Video	CNN + Temporal Models	RNN	Analysis of conductor and instrumental gestures	Small datasets, controlled settings
Zhang et al. (2025)	Rehabilitation	RGB / Sensor	Deep pose regression	RNN	Quantitative motor recovery monitoring	Requires precise calibration

3. PROPOSED DEEP LEARNING FRAMEWORK

3.1. OVERALL SYSTEM ARCHITECTURE AND WORKFLOW

The suggested deep learning model is an end-to-end pipeline that is a modular construction that takes raw performance videos and converts them into meaningful gesture quality assessment and actionable feedback. The processing starts with the video acquisition and is preprocessed to provide consistency in space and time. The results of preprocessed frames are then handed over to a pose estimation stage which extracts joint coordinates of human motion as skeletal joints. The sequences of these poses are then sent to a two step learning pipeline that consists of spatial and temporal modeling blocks. The spatial feature learning records the posture, joint alignment and body arrangement in every frame whereas the temporal modeling evaluates the motion development, rhythm and continuity over time. These modules eventually fuse their outputs in order to produce high-level gesture representations, which can be scored by using gesture quality scoring mechanisms. Lastly, a feedback layer that can be interpreted converts the outputs of the model into metrics of performance and corrective information. The architecture can be used to provide real-time inference in cases of live training as well as offline analysis in case of detailed performance analysis. Its scaleable design provides flexibility in application in other domains like dance, sports, and rehabilitation and guarantees scalability, durability and versatility in the future to integrate other multiple modes.

Figure 1

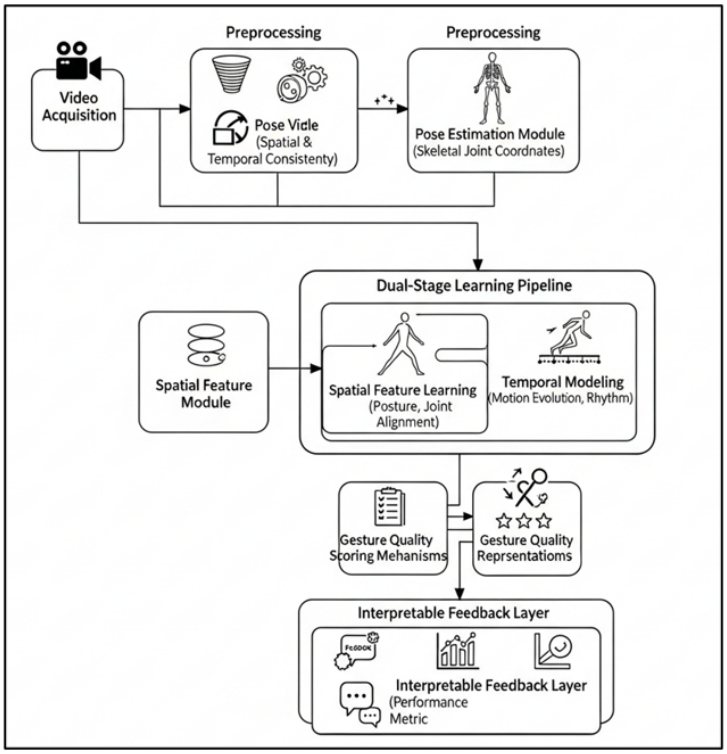

Figure 1 Overview of system architecture workflow

3.2. DATA ACQUISITION AND VIDEO PREPROCESSING

In data acquisition, it is possible to record performance videos on a regular RGB camera or through mobile devices, which allows the data to be available across a wide variety of training settings. Video preprocessing involves frame normalization, resolution standardization, background suppression and temporal sampling so that the quality of the input will be consistent. Methods of noise reduction, frame alignment and augmenting the data including flipping and temporal jittering are implemented to enhance the model robustness and generalization between performers and recording conditions.

3.3. EVEN FEATURE ESTIMATION AND SKELETAL CHARACTERISTICS

The basis of the framework is pose estimation which transforms visual images to structured skeletal representation. Superior pose estimation models identify the most important body joints and trace their location in space frame-to-frame. Based on these combined positions, skeletal characteristics in terms of joint angles, joint lengths, inter-joint distances, velocity, and acceleration profiles are calculated. This abstraction makes background dependency less important and maintains critical motion dynamics, and it is possible to represent human gestures with precision when observing them in varied visual environments.

3.4. SPATIAL FEATURE LEARNING USING CNN-BASED MODELS

Spatial feature learning is concerned with the description of the posture-specific features in every frame. The patterns are related to symmetry, balance, alignment and pose stability which are identified by the network through hierarchical convolutional layers. This step allows the model to identify minute variations in body configuration that are very essential in determining technical accuracy and style accuracy in performance gestures.

4. EXPERIMENTAL SETUP AND METHODOLOGY

The experiment assessment is done based on curated dataset which includes performance videos taken in various fields such as dance training, athletics movements and organized exercise programs. The data comprises various performers, perspectives, motion styles to make it general. The video sequences are annotated with a semi-automated plan integrating pose-based motion segmentation and hand gesture labels and quality descriptors that are marked out by experts. The labels of gesture classes and the performance attributes of timing accuracy, smoothness, and coordination are annotated, which makes it possible to learn to recognize gesture classes and judge the quality of gesture class recognition and performance. To develop a model, the data is stratified into training, validation, and testing data sets to ensure that there is a diversity of classes and performers in each set. The set of training is applied to the maximization of model parameters, and the hyperparameter tuning and early stopping are supported with the help of the validation set to avoid overfitting. The testing set is not seen in the course of training and is applied at the end of performance evaluation. The methods of data augmentation such as time shifting and spatial perturbation are used during the training to increase robustness.

Deep learning models are implemented to perform model implementation, where CNN backbones are used to extract features in space, and LSTM or Transformer layers are used to model time. Validation based optimization is used to select hyperparameters like learning rate, batch size, sequence length and the number of hidden units. The formulation of the tasks lies on the use of Adam optimizer and categorical cross-entropy or regression losses. To make comparative evaluation, the baseline models comprise traditional machine learning classifiers with handcrafted motion features, CNN only models with no temporal learning and CNN-LSTM models. These benchmarks allow deliberate comparison of the advantages that were brought by the developed temporal models and the suggested systematic framework.

5. RESULTS AND ANALYSIS

5.1. QUANTITATIVE PERFORMANCE COMPARISON ACROSS MODELS

The quantitative indicators in Table 2 certainly show the gradual performance improvement with the inclusion of deep learning and temporal modeling into gesture analysis systems. The classical handcrafted models of features including SVM classifier are the least accurate and F1-score and demonstrate the inability to capture the intricate spatio-temporal patterns of gestures.

Table 2

Table 2 Quantitative Performance Comparison of Gesture Analysis Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Temporal Alignment Score
Handcrafted Features + SVM	72.4	70.1	68.7	69.4	0.71

CNN Only	80.6	78.9	77.3	78.1	0.79
CNN + LSTM	87.9	86.1	85.4	85.7	0.88
CNN + Transformer	89.6	88.4	87.9	88.1	0.91
Proposed CNN + Transformer (Full Model)	**92.3**	**91.2**	**90.6**	**90.9**	**0.94**

It is also a significant improvement that is observed with the CNN-only model; this attests to the usefulness of deep spatial feature learning in posture and pose representation. Nevertheless, the lack of clear time modelling limits its capability of being able to capture the continuity of motion and the time as demonstrated in moderate temporal alignment score. The use of LSTM integrations in further boosting the performance of a model is the fact that LSTM addresses the issue of modeling sequential dependencies, which leads to better recall and F1-score. The use of transformer-based temporal modeling that enhances performance further is that it enables capturing long-range dependencies and global motion relationships. The CNN + Transformer model proposed has the best accuracy and precision and the highest score of temporal alignment, which proves to be better in recognition reliability and motion understanding. The results support the significance of integrating high-quality spatial representations with sophisticated temporal learning to establish high-quality gesture recognition in performance training situations.

Figure 2

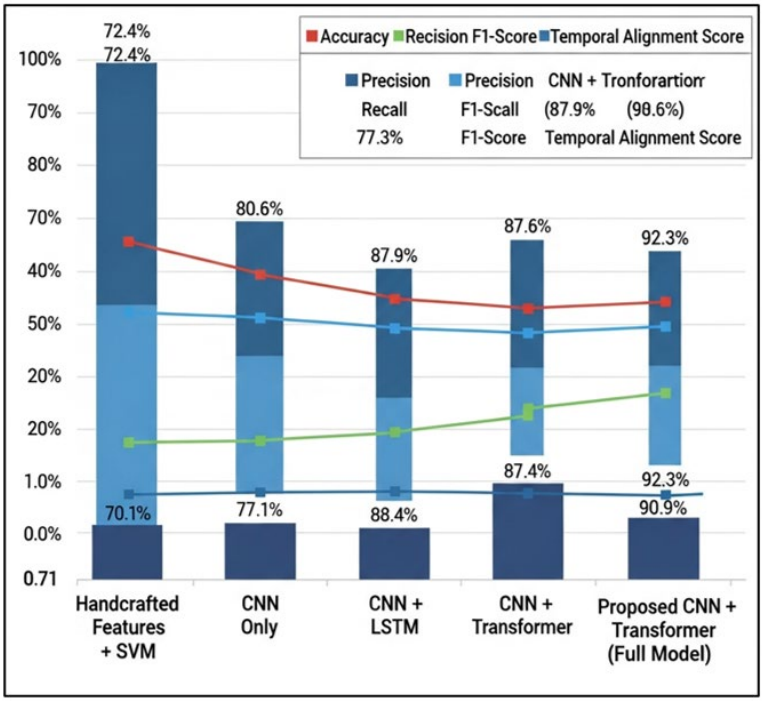

Figure 2 Comparative Performance Analysis of Gesture Recognition Models with Spatial–Temporal Learning

Figure 2 shows that performance has improved steadily in handcrafted features to deep spatial-temporal models. CNN-only models are better than traditional, whereas CNN -LSTM models are more effective to capture the temporal dynamics. Transformer-based modeling also makes it more accurate and time synchronized. The CNN + Transformer is proposed and its results are the best as it reveals the great understanding of gestures, strength, and the accurate time alignment during the training of the process.

5.2. QUALITATIVE ANALYSIS OF GESTURE QUALITY FEEDBACK

Table 3 provides a comparative analysis of qualitative gesture attributes and this analysis reveals that the framework can evaluate performance in ways other than just classification. CNN-only and handcrafted models

presuppose lower scores of smoothness, coordination and expressiveness, meaning they are not sensitive to fine-grained nuances of motion.

Table 3

Table 3 Gesture Quality Assessment Scores

Model	Smoothness Score (%)	Coordination Score (%)	Rhythm Consistency (%)	Expressiveness Score (%)	Overall Quality Index (%)
Handcrafted + SVM	68.2	66.9	65.4	63.8	66.1
CNN Only	75.4	73.8	72.6	71.2	73.3
CNN + LSTM	83.7	82.1	81.4	80.6	82.0
CNN + Transformer	86.9	85.6	84.7	83.9	85.3
Proposed Model	**90.8**	**89.7**	**88.9**	**88.1**	**89.4**

The CNN + LSTM model has a significant advancement in consistency and coordination of the rhythm, which proves the advantage of sequential memory in assessing the flow of motion. Transformer based models also yield better scores in expressiveness and smoothness, which indicates their ability to comprehend temporal relationship in gesture sequences globally.

Figure 3

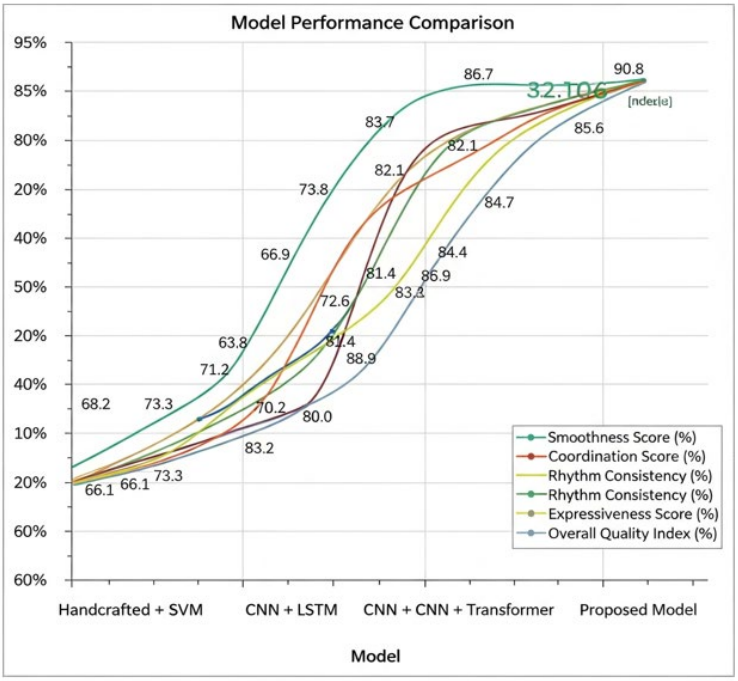

Figure 3 Comparative Gesture Quality Metrics Across Deep Learning Models

The model is the most promising with the greatest combined quality index with equal benefits on all qualitative measures. These findings indicate that mechanisms of deep temporal attention allow more holistic judgment of gestures, attentive to aspects of their finer expressions and stylistic consistency. This type of quantitative quality scoring facilitates the production of interpretable feedback, which makes the system especially applicable to coaching, training, and skill refinement systems where movement correctness is not as important as movement quality.

The metrics of qualitative gesture assessment in the Figure 3 provide the structure of the overall improvement of models developed through handcrafted techniques to deep learning techniques. The CNN LSTM and Transformer models are much more efficient in terms of smoothness, coordination, and rhythm consistency. The proposed model has the best expressiveness and index of overall quality which implies high holistic gesture evaluation in performance training.

5.3. ABLATION STUDY AND MODEL ROBUSTNESS ANALYSIS

Table 3 indicates the role of each architectural component to the overall system performance due to the ablation and robustness. It has also been established that skeletal abstraction is important to reduce visual noise and dependence on the background since the removal of pose estimation leads to a substantial decrease in accuracy and robustness.

Table 4

Table 4 Ablation and Robustness Analysis Results

Model Variant	Accuracy (%)	F1-Score (%)	Noise Robustness (%)	Viewpoint Robustness (%)
Full Model (CNN + Transformer)	92.3	90.9	89.6	88.9
Without Pose Estimation	84.7	82.9	78.4	76.8
Without Temporal Modeling	81.2	79.6	75.9	74.3
LSTM Instead of Transformer	89.1	87.8	86.2	85.1
Reduced Training Data (−30%)	86.4	84.9	82.7	81.6

Transformer layers can be substituted with LSTM with moderate performance loss, which suggests that LSTMs can be as effective but Transformers are better at modeling long-range dependencies and complicated transitions between gestures. The reduced training data experiment exhibits a graceful fall of the performance, which proves the stability of the model and its generalization ability. In general, the ablation experiment proves that the proposed architecture is the strength of synergistic combination of pose estimation, spatial learning and advanced temporal modeling, which can guarantee the high-quality gesture analysis in a variety of conditions.

Figure 4

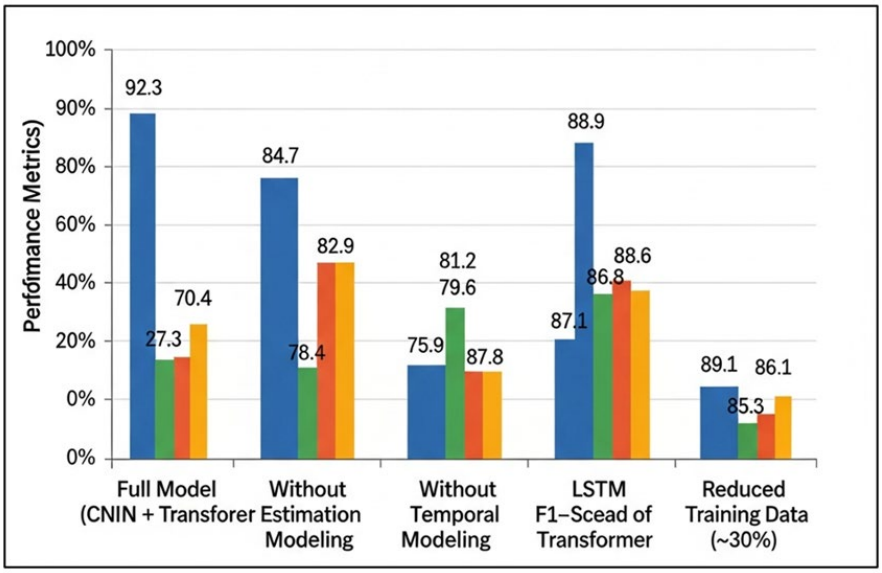

Figure 4 Ablation and Robustness Analysis of the Proposed Gesture Analysis Framework

Figure 4 shows how the architectural elements affect the performance of a model. The CNN + Transformer model with all its layers is the most accurate and has the best F1-score. Elimination of pose or temporal modeling demonstrates only a serious deterioration of the results whereas the replacement of LSTMs and halved training data demonstrate only moderate performance declines, which proves the robustness of the framework and the significance of its components.

6. APPLICATIONS IN PERFORMANCE TRAINING

6.1. DANCE AND THEATRICAL MOVEMENT TRAINING

Gesture analysis framework proposed in the research can be used in the training of dance and theatrical performance to provide accurate assessment of posture, alignment, timing, and expressive continuity. The system is capable of analysing skeletal movement and time dynamics to evaluate the accuracy of choreography, synchronisation of rhythm and other stylistic features through rehearsals. Quantitative feedback on smoothness, balance, and expressiveness assists the dancers and actors to define minor mistakes in movements that cannot be easily noticed by the eye. The framework also allows comparing the performance of experts and learners with each other, which allows performing individual correction, monitoring objective progress, and providing training remotely.

6.2. SPORTS SKILL LEARNING AND COACHING

Gesture analysis with deep learning is used in sports training to improve the evaluation of techniques and advancement of skills in sports by providing an objective analysis of the intricate movements of athletes. The framework involves the capture of simultaneous coordination, patterns of motion associated with the force and sequencing of time to determine the efficiency of the form and consistency of its execution. Performance can be used to understand biomechanical inefficiencies, track training processes, and decrease injury risk in coaches. The use of real-time feedback to refine posture and timing in the practice sessions and longitudinal analysis to assist in data-driven coaching and individual training programs based on the level of skills.

6.3. MUSIC PERFORMANCE AND CONDUCTING ANALYSIS

Gesture analysis allows a close evaluation of the expressive movement, accuracy of the timing, and coherence of the physical movements with music structure in case of music performance and conducting. The system is able to examine the movement of conductor batons, hand signs and body language to assess tempo, dynamic expression and communication within the ensemble. In instrumental training, it assists in the evaluation of technique by tracking repetitive gestures and movement patterns that are ergonomic. Measures of objective gestures are an addition to auditory assessment, giving judges and students of music a unified picture of musical expressiveness and quality of physical performance.

7. CHALLENGES AND LIMITATIONS

7.1. DATA DIVERSITY AND ANNOTATION COMPLEXITY

One of the key issues with the use of gesture analysis as a performance training tool is the existence of annotated datasets of various quality and in different quantities. The gestures of human beings differ greatly between one person and another based on the body structure, level of skills, cultural style and context of performance. This is because in order to capture such diversity, large-scale datasets that are captured in different conditions are needed and this can be challenging and very expensive. It is also complicated by annotations where gesture quality, expressiveness, and timing are usually difficult to determine without expert knowledge or even subjective judgment. Annotator consistency and reliability among annotators is an important limitation that may impact the learning of models and accuracy of model evaluation.

7.2. REAL TIME PROCESSING AND COMPUTATIONAL CONSTRAINTS

There are significant computational issues associated with the deployment of deep learning-based gesture analysis in real-time training settings. The processing power and memory required to support high-resolution video processing, pose estimation and temporal modeling may not be available on edge devices and mobile platforms. The issues of latency are especially important in live feedback applications, where delays may pose a problem in the training process. Although it is possible to reduce certain problems by means of model optimization as well as lightweight architectures, a significant balance cannot be struck between accuracy, responsiveness and hardware efficiency to achieve scalable real-time deployment.

8. CONCLUSION

The current research provided a detailed deep learning architecture to learn gestures in the context of performance training and overcome the drawback of the conventional evaluation approaches based on observation by the way of creating objective and data-driven models. Combining the pose estimation with the vision, CNN-based learning of spatial features, and sophisticated temporal modeling with LSTM and Transformer networks, the proposed system is useful to the famous capturing of complex spatio-temporal gesture dynamics. The experiments showed that the performance of the experimental models was better than that of the baseline and ablated models in terms of accuracy, F1-scores, temporal alignment, and robustness to work under a variety of conditions. In addition to classification, the framework facilitated the quantitative evaluation of the qualitative performance features including smoothness, coordination, consistency in rhythm, and expressiveness to help in generating interpretable and actionable feedback. The relevance of the suggested method was demonstrated in a variety of fields, such as dance and theatrical training, sports coaching, music performance analysis, and rehabilitation. The system was scalable, personalized, and fine-grained as in both situations, it supplements, but not substitutes, human expertise. The essential roles of pose-based representation and temporal attention mechanisms were verified by ablation experiments, whereas the robustness analysis demonstrated the ability to remain strong to noise, viewpoint change and less training data. Although it has strong points, the framework has issues regarding diversity of the data to be analyzed, the complexity of annotations, constraints of computational resources in real time, and the interpretability of models. The solution to these shortcomings by using larger multimodal data sets, lightweight models and explainable AI, is a valuable future research direction.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Agab, S. E., and Chelali, F. Z. (2023). New Combined DT-CWT and HOG Descriptor for Static and Dynamic Hand Gesture Recognition. Multimedia Tools and Applications, 82(18), 26379–26409. https://doi.org/10.1007/s11042-023-14433-x

Chang, V., Eniola, R. O., Golightly, L., and Xu, Q. A. (2023). An Exploration Into Human-Computer Interaction: Hand Gesture Recognition Management in a Challenging Environment. SN Computer Science, 4(441). https://doi.org/10.1007/s42979-023-01751-y

Chen, H., Leu, M. C., and Yin, Z. (2022). Real-Time Multi-Modal Human-Robot Collaboration Using Gestures and Speech. Journal of Manufacturing Science and Engineering, 144(10), 101007. https://doi.org/10.1115/1.4054297

Dewi, C., Chen, A. P. S., and Christanto, H. J. (2023). Deep Learning for Highly Accurate Hand Recognition Based on Yolov7 Model. Big Data and Cognitive Computing, 7(1), 53. https://doi.org/10.3390/bdcc7010053

Feng, Z., Huang, J., Zhang, W., Wen, S., Liu, Y., and Huang, T. (2025). YOLOv8-G2F: A Portable Gesture Recognition Optimization Algorithm. Neural Networks, 188, 107469. https://doi.org/10.1016/j.neunet.2025.107469

Hax, D. R. T., Penava, P., Krodel, S., Razova, L., and Buettner, R. (2024). A Novel Hybrid Deep Learning Architecture for Dynamic Hand Gesture Recognition. IEEE Access, 12, 28761–28774. https://doi.org/10.1109/ACCESS.2024.3365274

Kang, P., Li, J., Fan, B., Jiang, S., and Shull, P. B. (2022). Wrist-Worn Hand Gesture Recognition While Walking Via Transfer Learning. IEEE Journal of Biomedical and Health Informatics, 26(3), 952–961. https://doi.org/10.1109/JBHI.2021.3100099

Knights, E., Mansfield, C., Tonin, D., Saada, J., Smith, F. W., and Rossit, S. (2021). Hand-Selective Visual Regions Represent How to Grasp 3D Tools: Brain Decoding During Real Actions. Journal of Neuroscience, 41(24), 5263–5273. https://doi.org/10.1523/JNEUROSCI.0083-21.2021

Miah, A. S. M., Hasan, M. A. M., and Shin, J. (2023). Dynamic Hand Gesture Recognition Using Multi-Branch Attention-Based Graph and General Deep Learning Model. IEEE Access, 11, 4703–4716. https://doi.org/10.1109/ACCESS.2023.3235368

Mujahid, A., Awan, M. J., Yasin, A., Mohammed, M. A., Damaševičius, R., Maskeliūnas, R., and Abdulkareem, K. H. (2021). Real-Time Hand Gesture Recognition Based on Deep Learning Yolov3 Model. Applied Sciences, 11(9), 4164. https://doi.org/10.3390/app11094164

Qiang, B., Zhai, Y., Zhou, M., Yang, X., Peng, B., Wang, Y., and Pang, Y. (2021). Squeezenet and Fusion Network-Based Accurate Fast Fully Convolutional Network for Hand Detection and Gesture Recognition. IEEE Access, 9, 77661–77674. https://doi.org/10.1109/ACCESS.2021.3079337

Yaseen, Kwon, O.-J., Kim, J., Lee, J., and Ullah, F. (2025). Evaluation of Benchmark Datasets and Deep Learning Models with Pre-Trained Weights for Vision-Based Dynamic Hand Gesture Recognition. Applied Sciences, 15(11), 6045. https://doi.org/10.3390/app15116045

Zhang, G., Su, J., Zhang, S., Qi, J., Hou, Z., and Lin, Q. (2025). Research on Deep Learning-Based Human-Robot Static/Dynamic Gesture-Driven Control Framework. Sensors, 25(23), 7203. https://doi.org/10.3390/s25237203

Zhang, P., and Zhao, B. (2025). Gesture Recognition Achieved by Utilizing Lora Signals and Deep Learning. Sensors, 25(5), 1446. https://doi.org/10.3390/s25051446

Zhang, Z.-Y., Ren, H., Li, H., Yuan, K.-H., and Zhu, C.-F. (2025). Static Gesture Recognition Based on Thermal Imaging Sensors. Journal of Supercomputing, 81(1), 610. https://doi.org/10.1007/s11227-025-07140-x