

Transformer-Guided Monocular 3D Fitness Pose Analysis with Biomechanical Constraints and Privacy Preservation

¹Suchismita Mohapatra, ¹Mohit Ranjan Panda, ²Binod Kumar Pattanayak*

¹School of Computer Engineering Kalinga Institute of Industrial Technology (Deemed to be University)
Bhubaneswar, Odisha 751024, India Emails: 2081037@kiit.ac.in, mohit.pandafcs@kiit.ac.in,

²Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan Deemed to be University,
Bhubaneswar, Odisha 751030, India binodpattanayak@soa.ac.in

*Corresponding Author: Binod Kumar Pattanayak

Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan Deemed to be University,
Bhubaneswar, Odisha 751030, India binodpattanayak@soa.ac.in

HOW TO CITE:

Suchismita Mohapatra, Mohit
Ranjan Panda, Binod Kumar
Pattanayak (2026). Transformer-
Guided Monocular 3D Fitness
Pose Analysis with Biomechanical
Constraints and Privacy
Preservation.
International Journal of Special
Education, 41(13s), 953-981

ABSTRACT:

Monocular 3D pose estimation holds great promise in fitness and rehabilitation settings, but various challenges arise, including privacy concerns, interpretability limitations, and user and environment-dependent posture differences. The proposed approach is a comprehensive framework for fitness pose analysis, based on monocular RGB input, adopting the transformer architecture for 3D pose estimation, as well as a biomechanical classification and privacy-guarding output for the whole pipeline. The methodology includes image enhancement, reconstruction using a transformer, evaluation of the joint angles, watermarking and encryption of the data for postural assessment and secure data management. The framework was evaluated with a customised dataset of squat, plank and push-up exercises using correct and incorrect forms to determine the reliability of the system in performing real-life tasks. The proposed system was able to attain an accuracy of 72.73% with the implementation of XGBoost using a subject-independent 5-fold cross-validation strategy. This accuracy may seem average, but it is still under the controlled condition where there is a significant inter-subject variability. While the previous approaches relied on a controlled dataset and powerful deep learning classifiers, the proposed system places emphasis on the interpretability of these classifiers, computational efficiency, and practicality of deployment. Furthermore, it is shown that the framework can be applied to cross domains, specifically within the zero-shot evaluations of Human3.6M and MPI-INF-3DHP datasets, which is an important test, since no “re-training” operations are performed. This generalisation capability is further validated in a real-world, zero-shot setting via qualitative evaluations based on images not seen before that are publicly available. This shows the feasibility of developing such lightweight, easily-interpretable and privacy-preserving 3D fitness analysis systems without the need for specialised sensing devices.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Monocular 3D pose estimation, TransUNet, biomechanics, fitness monitoring, AES encryption, privacy-preserving AI

I. Introduction

Contemporary developments in computer vision and deep learning over the past few years have improved the accuracy of human pose estimation remarkably, which in turn opens up possibilities for various applications in the fitness monitoring, rehabilitation therapy and human-computer interaction systems [1]. Among these tasks, estimation of 3D human pose from a single monocular RGB image, which refers to the prediction of the 3D positions of certain human body parts from a single RGB image, has proven to be quite challenging because of inherent depth ambiguity, occlusion, and large variations in human body motion.

In traditional convolutional neural network (CNN) methods, the emphasis is primarily on learning localised features, and they are less

efficient in capturing the more holistic nature of the body joints themselves [2]. As a result, these can produce poses that are not necessarily biomechanically sound, particularly when doing exercises such as squats, pushups, and planks under analysis.

Beyond accurately reconstructing the image, fit-for-purpose fitness monitoring systems need to provide meaningful interpretability and implement robust data protection measures. The current approaches focus mainly on the accuracy of the pose estimation, with relatively less focus on biomechanical reasoning, safe handling of personal and sensitive user information, and assessment in free exercise settings.

However, to alleviate these limitations, this work introduces a transformer-based architecture that is physically constrained to the biomechanics of human fitness pose estimation, while also preserving privacy. The design of a unified transformer-optimisation framework that combines monocular 3D pose reconstruction, real-world interpretable biomechanical posture classification, and privacy-preserving secure analytics into a single, deployment-friendly

pipeline is the main contribution of this work. Unlike prior approaches that address these components independently, the proposed system enables simultaneous optimisation of accuracy, interpretability, and data security under real-world fitness conditions. To the best of my knowledge, this is one of the first works to jointly integrate transformer-based 3D pose estimation, interpretable biomechanical reasoning, and privacy-preserving mechanisms within a unified fitness-monitoring framework.

In contrast to large benchmark-driven studies, the dataset represents a subject-diverse real-world fitness benchmark consisting of 110 participant-level recordings collected under unconstrained conditions, designed to evaluate practical deployment scenarios rather than controlled laboratory generalisations. Subject-wise grouping across squat, plank, and push-up exercises, the benchmark captures realistic home and gym exercise settings with inter-subject variability, posture deviations, and environmental noise. In addition, qualitative validation on unseen real-world rain imagery provides preliminary evidence of generalisation beyond the collected dataset. Unlike existing works that address pose estimation, biomechanical analysis, and privacy separately, this work presents a unified deployment-oriented framework that integrates all three components into a single pipeline for real-world fitness monitoring.

Contributions

The main contributions of this paper are as follows:

- A unified transformer-guided pipeline that jointly integrates monocular 3D pose reconstruction, interpretable biomechanical posture classification, and privacy-preserving secure analytics.
- A hybrid framework combining transformer-based 3D pose estimation with interpretable biomechanical feature extraction and lightweight machine learning classification.

- Interpretable biomechanical feature extraction for exercise quality evaluation.
- Construction and evaluation of a real-world participant-level fitness dataset containing 110 recordings across three exercises with balanced correct/incorrect posture labels under subject-disjoint evaluation protocols.
- Zero-shot cross-domain generalisation analysis on Human3.6M and MPI-INF-3DHP without retraining.
- A practical privacy-preserving framework integrating encryption and watermarking.

Organisation of the Paper

The remainder of this paper is organised as follows. Section II reviews related work. Section III presents the proposed methodology. Section IV describes the experimental setup. Section V reports the results and analysis. Section VI discusses the findings and limitations. Section VII concludes the paper and outlines directions for future work.

II. Related Work

Recent progress in human pose estimation has been driven by advances in deep learning, large-scale datasets, and transformer-based architectures. Existing studies may be broadly grouped into four categories: (A) monocular 3D human pose estimation, (B) transformer-based vision architectures, (C) fitness and posture assessment systems, and (D) privacy-preserving intelligent systems. This section reviews these directions and positions the proposed framework within the current literature.

A. Monocular 3D Human Pose Estimation

Monocular three-dimensional human pose estimation involves the reconstruction of joint coordinates in three-dimensional space from individual RGB images or video sequences. As compared to multi-camera motion capture techniques and depth-sensing methods, this method is more practical and scalable for various applications, such as fitness tracking, rehabilitation, and sports analysis. However, the method still had many problems to overcome because of the inherent ambiguity of depth, the

phenomenon of self-occlusion, the perspective of the viewpoints, the background clutter, and so on, as described by Liu et al [1].

The fundamental structured-model-based approaches laid down important groundwork for the current methods used for pose estimation [3]. Later, regression techniques were used to build a convolutional model that was trained to extract the three-dimensional coordinates of the joints from visual input or the heatmap representation. Martinez et al. [4] showed that a highly accurate three-dimensional pose reconstruction could be obtained using simple lifting networks based only on two-dimensional keypoint data. To build on this, Pavlo et al. [5], and collaborators used temporal consistency patterns in video sequences to increase system reliability.

Modern transformer-based methods have improved the performance by enhancing their ability to extract long-range spatial and temporal relationships over long distances. Zheng et al. [6] introduced PoseFormer; Zhao et al. [7] improved efficiency with PoseFormerV2; Zhang et al. [8] proposed MixSTE for mixed spatio-temporal encoding; and Li et al. [9] introduced HourglassTokeniser to improve transformer efficiency for video-based 3D pose estimation. Graph-transformer approaches [10] further incorporate skeletal topology for more structurally consistent reconstruction.

Despite recent advances in this area, most current studies emphasise benchmark performance instead of biomechanical interpretability, efficient deployment, and privacy protection measures. Current studies in sports applications have shown that models built based on typical datasets suffer in performance for high-acceleration sport movements, thereby underscoring the importance of specific assessment protocols for sport and fitness applications [11]. In response to these identified limitations, the proposed framework is designed to combine monocular 3D pose estimation and posture analysis, while security protocols are embedded to facilitate intelligent monitoring of fitness.

B. Transformer-Based Vision Architectures

Transformer-based vision models have become popular for their strong performance in capturing long-range relationships and global context information from visual data, as opposed to conventional CNN-based methods. Dosovitskiy et al. [2] introduced the Vision Transformer (ViT), while Liu et al. [12] improved practicality with the Swin Transformer through shifted-window attention and multi-scale feature learning.

Hybrid architectures combine local convolutional priors with global self-attention. Chen et al. [13] showed that TransUNet improves segmentation quality by integrating CNN feature extraction with transformer encoders, and Hatamizadeh et al. [14] further demonstrated the suitability of transformer backbones for structured volumetric reconstruction through UNETR.

Application-driven studies have reinforced this generalizability. Improved TransUNet variants [15] have enhanced reconstruction in inverse imaging tasks such as electromagnetic tomography, while Hassan et al. [16] showed that attention-driven DeiT-based models are effective for dynamic human activity recognition.

The results show that transformer-based architectures are effective when used in posture estimation systems, as they provide the benefit of combining high-level structural information with high-resolution anatomical accuracy while maintaining computational efficiency—an important factor for practical applications [12].

C. Fitness and Posture Assessment Systems

As people are increasingly looking for digital fitness solutions and personalised health care services, the evolution of intelligent posture evaluation systems that leverage visual or wearable sensor-based exercise performance evaluation has been propelled. As opposed to manual supervision techniques, these automated solutions provide scalable, objective and real-time feedback systems which are ideal for home-

based exercise programs, rehabilitation, and athletic training applications. Assessment of posture is well challenged, though, as it must account for differences in the observer's viewpoint, partial occlusion of body segments, inter-individual anatomical variations, and the nuances of good and bad movement execution.

Recent vision-based studies have demonstrated the benefits of pose estimation to the field of exercise assessment. Kahl et al. [17] investigated pose estimation methods using video cameras and found high agreement between video-based methods (e.g., MeTRAbs, AlphaPose, MediaPipe, PoseFormerV2) and motion capture in exercise scenarios. Rao et al. [18] introduced a real-time squat posture correction system that classified the current squat posture into a set of categories and provided corrective feedback to the user through a Bi-CGRU model based on MediaPipe landmarks.

Multimodal coaching systems further extend this direction. Yin et al. [19] introduced MRCoach, which combines IoT and mixed reality to reconstruct user motion and deliver personalised low-latency feedback. Tang et al. [20] showed that smart sportswear based on textile strain sensors can continuously assess movement symmetry, breathing coordination, and biomechanical performance. Liu et al. [21] further improved motion representation and activity recognition through affinity contrastive learning on structured skeleton sequences.

Although significant progress has been made in this field, the majority of current systems focus exclusively on individual components such as posture detection, wearable sensor technology, or activity categorisation. These methods are not sufficient to meet the demand for a complete mechanism to protect privacy while allowing the data to be utilised for interpretable skeletal analysis, 3D structural evaluation, and strong data security protocols [22]. Based on these identified limitations, this research introduces an all-in-one framework comprising of a combination of 3D reconstruction methods, exercise posture classification algorithms, and integrated security

protocols to ensure trustworthy fitness monitoring systems.

D. Privacy-Preserving Intelligent Systems

As intelligent vision technologies are being employed in a variety of healthcare, fitness, and smart environment applications, there has been growing concern over privacy protection, ownership rights, and the secure implementation of these technology solutions. Consider the nature of the visual information, and that this could reveal individual identity, physical features and behavioural patterns, it is important that there are strong protections for confidentiality, authenticity of the data, and against unauthorised intrusion in any practical posture evaluation systems. The encryption of images is an essential technique for secure storage and transmission. Alghamdi et al. [23] summarized the image encryption techniques and highlighted the ongoing relevance of hybrid, RSA, and AES-based images in the context of secure multimedia and medical applications. Deebak et al. [24] also demonstrated that lightweight encryption can be used to allow privacy-preserving visual analytics with manageable overhead on industrial IoT devices. Digital watermarking adds ownership verification and content authentication. Singh et

al. [22] combined encryption and dual watermarking for secure healthcare data exchange, while Zhao et al. [25] and Yuan et al. [26] extended watermarking to AI models for tampering detection and authentication.

Despite these advances, security is often treated separately from posture recognition. The proposed framework addresses this void by leveraging a combination of AES-based protection and watermarking-based authenticity preservation for trustworthy and privacy-aware fitness monitoring.

To clarify the research gap more explicitly, Table I compares representative prior studies across reconstruction performance and practical deployment capabilities. Summarised in Table I, benchmark-focused 3D pose estimation methods mainly optimise MPJPE, whereas fitness monitoring systems emphasise feedback and usability. In contrast, the proposed framework is designed to unify monocular 3D pose estimation, posture classification, biomechanical interpretability, practical deployment usability, and security within one pipeline. In the table, the symbols ✓, ×, and △ denote supported, not supported, and partially supported capabilities, respectively.

Table I: Diagnostic Comparison of Representative Related Work

Method	Yr	Data	Result	3D	Cls.	Int.	RT	Sec.
Martinez et al. [4]	2017	H3.6M	37.1 mm MPJPE	✓	×	×	×	×
Pavlo et al. [5]	2019	H3.6M	46.8 mm MPJPE	✓	×	×	×	×
PoseFormer [6]	2021	H3.6M	44.3 mm MPJPE*	✓	×	×	×	×
MixSTE [8]	2022	H3.6M	43.8 mm MPJPE*	✓	×	×	×	×
Rao et al. [18]	2025	Squat set	Real-time correction	×	✓	✓	✓	×
MRCoach [19]	2025	Sports data	79.9 ms, 56+ FPS	△	✓	✓	✓	△
Smart sportswear [20]	2025	Sensor data	Continuous monitoring	×	✓	✓	✓	△
Proposed	2026	110 participant-level recordings	92.17 mm MPJPE (LOSO)	✓	✓	✓	△	✓

The diagnostic comparison in Table I suggests that the proposed framework is not a straight replacement for a Human3.6M optimised method. Rather, its addition is its support of functional integration in realistic fitness monitoring scenarios with 110 participant recordings using a leave-one-subject-out cross-validation (LOSO), coupled with the maintenance of 3D structural reasoning and privacy protection, which are usually lacking in previous exercise-analysis systems.

III. Methodology

The framework presented establishes a transformer-based, privacy-conscious methodology for estimating three-dimensional human poses from single-camera inputs within fitness monitoring contexts. This system simultaneously pursues three core goals: precise reconstruction of 3D poses, comprehensible classification of postural configurations, and protected management of generated data outputs. The overall architecture is illustrated in Fig 1.

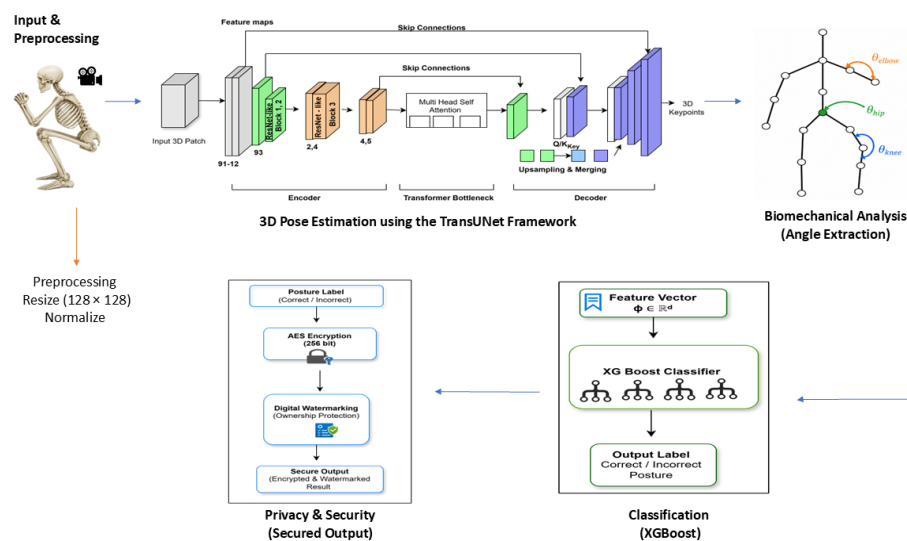


Fig. 1. Overview of the proposed transformer-guided framework for monocular 3D fitness pose analysis. The pipeline combines the capabilities of 3D pose estimation (using TransUNet), biomechanical feature extraction, posture classification (using XGBoost), and privacy-preserving secure output generation.

The proposed framework comprises three primary modules: (i) 3D pose estimation utilising TransUNet architecture, (ii) extraction of biomechanical features for comprehensive posture analysis, and (iii) AES-based encryption mechanisms for generating secure outputs. This system processes RGB video data as input and generates 3D joint coordinates alongside posture classifications and encrypted outputs. The operational sequence encompasses frame selection and preprocessing stages, followed by 3D pose estimation procedures, joint-angle computation, and posture categorisation, and culminates in secure output generation through integrated encryption and watermarking techniques.

A. Problem Statement

This work is aimed at developing a reliable and explainable monocular 3D human pose estimation system with the sole use of RGB video information for fitness monitoring purposes. Given an individual image or given a video frame that contains a person performing an exercise movement, the system tries to predict the 3D joint coordinates of the person, compute biomechanical metrics such as joint angles, and determine the correctness of the observed postural configuration.

This problem is challenging because monocular 3D pose estimation is affected by depth ambiguity, self-occlusion, background clutter,

viewpoint variation, and illumination changes [4]. In addition, practical fitness monitoring systems must provide explanations that are understandable to users while securely handling potentially sensitive output data.

The aim of the proposed methodology is the development of an integrated framework that simultaneously meets three important requirements: accurate 3D pose fulfillment, clear posture assessment using biomechanical analyses, and safe output handling by using encryption and watermarking techniques, ensuring privacy.

Formally, let's $I \in \mathbb{R}^{H \times W \times 3}$ denote an input RGB frame. The objective is to learn the mapping defined in Eq (1):

$$\mathcal{F}: I \rightarrow \hat{P}_{3D}, \quad (1)$$

where $\hat{P}_{3D} \in \mathbb{R}^{J \times 3}$ represents the predicted 3D coordinates of J body joints. From $\hat{P}_{3D}$, a biomechanical feature vector $\hat{b}$ and posture label $\hat{y}$ are obtained as in Eq (2):

$$\hat{b} = \phi(\hat{P}_{3D}), \quad \hat{y} = f_{XGB}(\hat{b}), \quad (2)$$

where $\phi(\cdot)$ denotes joint-angle extraction and $f_{XGB}(\cdot)$ denotes XGBoost-based posture classification. Adapted from prior TransUNet-based pose estimation and pose-driven biomechanical assessment formulations [13], the overall inference pipeline may therefore be summarized compactly as

$$\hat{y} = f_{XGB}(\phi(f_{TransUNet}(I; \Theta))), \quad (3)$$

which highlights the end-to-end flow from RGB input to 3D pose reconstruction, biomechanical feature extraction, and final posture assessment. This compact composition is an adapted framework-level formulation based on the TransUNet pose estimation model and pose-based biomechanical assessment pipeline described in prior work.

B. TransUNet-Based 3D Pose Estimation

The 3D pose estimation module is based on a TransUNet architecture that combines convolutional neural networks with transformer-

based global attention. TransUNet is selected due to its ability to combine local spatial precision (CNN) with global contextual reasoning (transformer), making it suitable for structured human pose reconstruction under noisy real-world conditions. The encoder extracts hierarchical spatial features from input RGB frames, capturing local anatomical details such as joints and limbs.

To overcome the limitations of purely convolutional models, a transformer module is incorporated at the bottleneck. This module utilises multi-head self-attention to model long-range dependencies between body joints, enabling improved global structural understanding.

The decoder reconstructs spatial features through progressive upsampling and generates 3D joint coordinates. Skip connections between encoder and decoder layers preserve fine-grained spatial information, ensuring accurate estimation of joint positions and body structure.

Given an input frame I , the network predicts the 3D pose as given in Eq (4), following the TransUNet-style encoder-transformer-decoder formulation in Chen et al. [13]:

$$\hat{P}_{3D} = f_{TransUNet}(I; \Theta), \quad (4)$$

where Θ denotes the learnable model parameters. The pose estimation objective is defined by the mean squared error in Eq (5), consistent with standard supervised 3D pose regression practice in prior work [4]:

$$\mathcal{L}_{pose} = \frac{1}{J} \sum_{j=1}^J \|\hat{p}_j - p_j\|_2^2, \quad (5)$$

where p_j and $\hat{p}_j$ denote the reference and predicted 3D coordinates of the j th joint, respectively.

This hybrid architecture enables the model to capture both local and global relationships, resulting in more stable and biomechanically consistent pose predictions. Together with the biomechanical mapping in Eq (3), this formulation clarifies how the reconstructed 3D

pose is transformed into interpretable posture labels within the proposed framework.

C. Biomechanical Feature Extraction

The proposed framework not only reconstructs the pose in 3D but also utilises biomechanical feature extraction to facilitate the analysis of posture in an interpretable manner. Using the 3D joint positions predicted with 3D joint coordinates, important joint angles are derived based on the geometric relationship between body joints.

Specifically, the following joint angles are extracted:

- Knee angle, computed using the hip-knee-ankle joint configuration.
- Hip angle, computed using the shoulder-hip-knee joint configuration.
- Elbow angle, computed using the shoulder-elbow-wrist joint configuration.

These joint angles are used as biomechanical indicators of the correctness of the posture. A correct squat has the correct knee and hip bend, and the incorrect posture will have angles that vary from what is expected.

The extracted features are used to classify the posture as correct or incorrect so that the posture is evaluated in an interpretable and explainable manner. This solves the problem of deep learning predictions and human-understandable biomechanics.

Let the predicted 3D pose be represented as in Eq (6), using the standard joint-set notation adopted in monocular 3D pose estimation literature [4]:

$$\hat{\mathbf{P}}_{3D} = \{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \dots, \hat{\mathbf{p}}_J\}, \quad \hat{\mathbf{p}}_j = (x_j, y_j, z_j), \quad (6)$$

where $\hat{\mathbf{p}}_j$ denotes the 3D coordinate of the j th body joint.

For any three joints A, B, and C, the corresponding limb vectors are defined in Eq (7):

$$\mathbf{u} = \mathbf{A} - \mathbf{B}, \quad \mathbf{v} = \mathbf{C} - \mathbf{B}. \quad (7)$$

The joint angle formed by these three points is then computed using the standard dot-product formulation in Eq (8), which is widely used in pose-based posture analysis:

$$\theta = \cos^{-1} \left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \right). \quad (8)$$

Using this formulation, the major biomechanical features are obtained through the knee, hip, and elbow angle definitions in Eq (9), Eq (10), and Eq (11), following the posture-analysis setting in Rao et al. [18]:

$$\theta_{\text{knee}} = f(\hat{\mathbf{p}}_{\text{hip}}, \hat{\mathbf{p}}_{\text{knee}}, \hat{\mathbf{p}}_{\text{ankle}}), \quad (9)$$

$$\theta_{\text{hip}} = f(\hat{\mathbf{p}}_{\text{shoulder}}, \hat{\mathbf{p}}_{\text{hip}}, \hat{\mathbf{p}}_{\text{knee}}), \quad (10)$$

$$\theta_{\text{elbow}} = f(\hat{\mathbf{p}}_{\text{shoulder}}, \hat{\mathbf{p}}_{\text{elbow}}, \hat{\mathbf{p}}_{\text{wrist}}), \quad (11)$$

where $f(\cdot)$ denotes the angle computation defined in Eq (8). The resulting biomechanical descriptor is then expressed in Eq (12):

$$\hat{\mathbf{b}} = [\theta_{\text{knee}}, \theta_{\text{hip}}, \theta_{\text{elbow}}]^T. \quad (12)$$

This formulation is applied to the relevant joint triplets for knee, hip, and elbow angle estimation.

D. Frame Selection and Preprocessing

To ensure stable and meaningful evaluation, representative frames are selected from each input video. Instead of using arbitrary frames, the middle frame of each video sequence is chosen, as it typically corresponds to the most stable posture and avoids noise present at the beginning and end of motion sequences.

Each selected frame is resized to a resolution of 128×128 pixels to reduce computational complexity while preserving essential spatial features. In order to enhance the visual clarity of the image, various image enhancement procedures such as brightness and contrast adjustments are applied. The normalized input frame is defined as Eq (13):

$$I_{\text{norm}} = \frac{I - \mu}{\sigma}, \quad (13)$$

where μ and σ denote the mean and standard deviation used for channel-wise normalisation. This preprocessing step ensures that the input frames are suitable for both pose estimation and visualisation.

E. Posture Classification Strategy

The Transformer-based TransUNet model is employed for robust 3D pose estimation, while the classification stage utilises interpretable biomechanical features rather than end-to-end deep learning, enabling transparency and reduced computational complexity.

Posture classification is performed using a supervised machine learning approach based on biomechanical features. Specifically, an XGBoost classifier is employed to learn the decision boundary between correct and incorrect postures.

The extracted joint-angle features (knee, hip, elbow) form a low-dimensional, interpretable feature space. The XGBoost classifier is trained using these features under subject-independent 5-fold cross-validation. Formally, the classification function is defined in Eq (14) as

$$\hat{y} = f_{XGB}(\mathbf{b}), \quad (14)$$

where $\mathbf{b} = [\theta_k, \theta_h, \theta_e]^T$ represents the biomechanical feature vector and $f_{XGB}(\cdot)$ denotes the trained XGBoost model.

This method provides robustness in classification over rule-based thresholding and also maintains the interpretability of the classification using the biomechanical features.

F. Algorithm

The general algorithm of the proposed system is summarized in Algorithm 1. The algorithm is implemented in a single pipeline, including preprocessing, TransUNet-based 3D pose estimation, biomechanical feature extraction, posture classification using XGBoost, and secure output generation.

Algorithm 1 Transformer-Guided Privacy-Preserving Fitness Monitoring Framework

Require: Input video V

Ensure : Predicted 3D pose $\hat{P}_{3D}$, biomechanical descriptor $\hat{\mathbf{b}}$, posture label $\hat{y}$, encrypted output C

1. Select a representative frame I from V
2. Resize I to 128×128 and normalize it to obtain I_{norm}
3. Estimate 3D joint coordinates using TransUNet: $\hat{P}_{3D} \leftarrow f_{TransUNet}(I_{norm}; \theta)$
4. Compute joint angles θ_{knee} , θ_{hip} , and θ_{elbow}
5. Form the biomechanical descriptor $\hat{\mathbf{b}} \leftarrow [\theta_{knee}, \theta_{hip}, \theta_{elbow}]^T$
6. Determine posture label $\hat{y} \leftarrow f_{XGB}(\hat{\mathbf{b}})$
7. Encrypt the posture label using AES: $C \leftarrow E_K(\hat{y})$
8. Apply watermarking to visualisation outputs for ownership protection
9. **return** $\hat{P}_{3D}$, $\hat{\mathbf{b}}$, $\hat{y}$, and C

G. Security Integration

Ensuring data privacy and protecting the ownership of fitness data have become critical requirements as the number of fitness monitoring systems based on AI grows; these systems must prioritise anonymity and security to safeguard personal information, for instance, through dual watermarking techniques [23]. The proposed system adopts a dual-layer security mechanism; it is based on the AES algorithm to achieve confidentiality, and a visual watermark is employed to achieve data integrity [27].

1. AES-Based Encryption

The classification result from the posture analysis module is protected by Advanced Encryption Standard (AES), which ensures the security of sensitive user information. The system implemented in the system encrypts the predicted posture label (correct or incorrect) before storing or visualising it.

Encryption helps protect user information during data transfer and storage by ensuring that it cannot be read without the correct decryption key. If m denotes the posture label and K denotes

the secret key, the encryption and decryption steps are represented conceptually in Eq (15) based on the AES standard:

$$C = E_K(m), \quad m = D_K(C), \quad (15)$$

where $E_K(\cdot)$ and $D_K(\cdot)$ denote AES encryption and decryption, respectively. In the visualisation stage, the encrypted label is decrypted to check the correctness and integrity of the encryption process. This shows that secure inference can be easily pipelined without compromising the performance of the system.

2. Watermarking for Data Ownership

In addition to encryption, visual watermarking is incorporated to provide ownership identification and prevent unauthorised reuse of fitness data. A watermark is embedded directly into the video frames in the spatial domain, ensuring visibility while preserving the quality of the input and reconstructed outputs.

The watermark serves as a lightweight mechanism for protecting intellectual property and enhancing trust in AI-generated outputs. In the implemented workflow, watermarking is applied to frames used for visualisation and ownership marking, while AES encryption is applied to the generated posture labels before storage or transmission. A simplified watermark embedding model is expressed in Eq (16), adapted from common visible watermarking formulations used in secure media pipelines [25],

$$I_w = (1 - \alpha)I + \alpha W, \quad (16)$$

where I denotes the original visualization frame, W denotes the watermark pattern, I_w denotes the watermarked frame, and α controls the embedding strength. It does not set the estimation or biomechanical analysis of the pose and can be deployed with a short latency.

3. Secure and Interpretable Framework

The AES-based encryption and watermarking offer a comprehensive security solution. Encryption outputs the classification information, while watermarking outputs the information on traceability and ownership. The

proposed framework incorporates these mechanisms into the pose estimation process, ensuring a balance of performance, interpretability, and security for practical applications in fitness monitoring.

IV. Experiments

This section introduces the experimental system, the description of the datasets, and the evaluation metrics employed to evaluate the performance of the proposed framework. The experiments are aimed at assessing the 3D pose reconstruction accuracy and, at the same time, the effectiveness of posture classification in a real-life environment.

A. Dataset Description

This data set contains 110 exercise recordings from 110 different participants, which means that each sample represents a separate exercise recording from a different person. There are several exercise categories, including plank, push-up, and squat, and the correct and incorrect forms. This in-between multi-tasking, multi-modal dataset allows for a comprehensive assessment and generalisation of performance over unseen people.

1. Participant Diversity

The data covers a diverse range of body sizes, genders, levels of exercise experience, clothing types, and recording settings, enhancing the realism of the evaluations. In order to ensure a fair and less biased sample, the number of male and female subjects is balanced in the sample. This balance enables dependable biomechanical analysis for various body structures and enhances the overall generalisation of the framework.

2. Environmental Variation

The dataset is composed to simulate in real-life scenarios, changes in camera perspective, lighting conditions, and subject motion patterns in order to mimic practical fitness monitoring situations. This variability enables the proposed system to be evaluated in realistic conditions, not under carefully controlled laboratory conditions. Videos were recorded with different indoor lighting, gym backgrounds, distances between the

camera and the participants, and side/front viewpoints and different smartphone devices.

3. Annotation Protocol

Each recording was manually verified and labelled as correct or incorrect based on observable biomechanical criteria such as

alignment, range of motion, and stability. The distribution of participants across exercise types and gender is presented in Table II. The dataset contains 54 correct and 56 incorrect participant samples, providing a nearly balanced class distribution.

Table II : Exercise and Gender- Wise Distribution Of The Participant Dataset

Exercise Type	Gender	Correct (PASS)	Incorrect (FAIL)	Total Samples
Squat	Male	8	8	16
Squat	Female	8	8	16
Plank	Male	7	8	15
Plank	Female	8	8	16
Push-up	Male	11	12	23
Push-up	Female	12	12	24
Total	Combined	54	56	110

4. Dataset Challenge Statement

As a moderately realistic dataset for participant-level fitness posture research involving manual collection, exercise execution diversity, and subject-wise evaluation, this dataset was intentionally designed to capture realistic posture variability, viewpoint changes, and execution errors commonly encountered in practical home fitness environments. The dataset, therefore, serves as a real-world validation set rather than a large controlled benchmark. Compared with laboratory datasets, the proposed dataset introduces uncontrollable artifacts, occlusions, and posture ambiguity, making classification more difficult but more representative of deployment conditions.

The dataset emphasises real-world applicability, where large-scale annotated 3D fitness datasets are often unavailable, thereby providing a meaningful testbed for practical deployment. In addition to the proposed fitness cross-domain evaluation, there are two public benchmarks: MPI-INF-3DHP and Human3.6M. MPI-INF-

3DHP provides diverse indoor and outdoor motion sequences with substantial viewpoint variation, whereas Human3.6M offers large-scale laboratory motion-capture recordings under more controlled conditions [28]. For the Human3.6M-based external evaluation in this study, RGB and 3D pose data are obtained from the H3WB benchmark built on Human3.6M [29]. These datasets are used only for generalisation and are not part of the primary Leave-One-Subject-Out Cross-Validation (LOSO) protocol on the proposed 110-participant fitness dataset.

B. Training Strategy

Pseudo ground truth is generated using MediaPipe pose estimation [30]. These pseudo-annotations provide approximate 3D joint coordinates that are used to supervise the training of the TransUNet-based model.

The model is implemented in PyTorch and trained using RGB frames resized to 128×128 pixels. All experiments were conducted in Python using PyTorch and scikit-learn on Google Colab with NVIDIA T4 GPU support. Training

focuses on learning the mapping between 2D image features and corresponding 3D joint structures. The model was optimised using Adam [31] with mini-batch training, and hyperparameters were selected empirically. To evaluate subject-independent performance, Leave-One-Subject-Out (LOSO) cross-validation was employed. Since each participant contributes exactly one sample, LOSO evaluates the model across 110 independent folds, where each fold tests on a completely unseen individual. The protocol assures that there is no subject overlap between training and testing and, hence,

a meaningful evaluation of generalisation performance.

Biomechanical features are computed from the predicted joint coordinates, in addition to pose estimation, to provide information for posture classification. Such deep learning and geometric analysis work together to facilitate an interpretable fitness assessment. The loss curves for training and validation of the TransUNet model show stable loss values during optimisation and consistent loss behaviour throughout epochs, as displayed in Figure 2

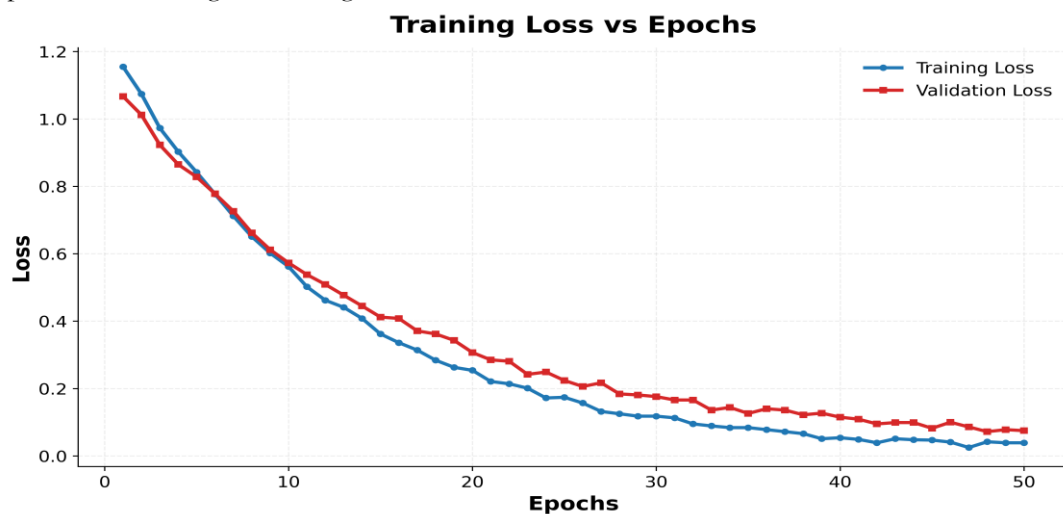


Fig.2. The training and validation loss curves for the proposed model, using TransUNet, converged stably at 50 epochs with a minimal amount of overfitting while showing some signs of generalisation.

Training and validation loss both go down steadily throughout the optimisation process, as seen in Fig. 2. This small difference between the curves suggests that proper generalisation with no significant overfitting is achieved, which helps validate the proposed framework for real-world fitness pose analysis.

C. Evaluation Metrics

The proposed framework is evaluated using both quantitative and qualitative criteria.

1. **3D Pose Estimation Metric:** The accuracy of 3D pose reconstruction is evaluated using Mean Per Joint Position Error (MPJPE), defined in Eq (17), which is the standard metric used in monocular 3D pose estimation [4] and measures the average Euclidean distance between predicted and reference joint positions:

$$MPJPE = \frac{1}{J} \sum_{j=1}^J \|\hat{p}_j - p_j\|_2. \quad (17)$$

2. **Biomechanical Analysis:** Joint angles, including knee, hip, and elbow angles, are computed to assess posture correctness. These angles can give helpful information about how well the exercise was performed.

3. **Visualisation-Based Evaluation:** A visualisation pipeline is developed to qualitatively measure the performance of the model. The output consists of the input frame, the reconstructed 3D pose, joint angle annotations, and classification results for each output.

4. **Security Evaluation:** AES-based encryption is proven to be effective, and the

classification output is encrypted and decrypted during inference to handle user data securely.

For posture classification, the standard evaluation measures are defined as follows. The classification metrics are given in Eq (18), Eq (19), and Eq (20), while the subject-independent accuracy summary is given in Eq (21).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN'} \quad (18)$$

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN'} \quad (19)$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (20)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively. For the subject-independent evaluation protocol, the mean classification accuracy is computed as

$$Accuracy_{LOSO} = \frac{1}{S} \sum_{s=1}^S \mathbf{1}(\hat{y}_s = y_s), \quad (21)$$

where S denotes the total number of held-out subject samples, y_s is the ground-truth posture label for sample s , $\hat{y}_s$ is the predicted label, and $\mathbf{1}(\cdot)$ is the indicator function.

D. Implementation Details

The proposed framework is realised in Python and PyTorch. Experiments are run in the GPU-enabled environment to speed up model inference. The input frames are scaled to 128×128 pixels and normalised prior to the TransUNet model processing.

Representative frames are sampled from each video for visualisation, and 3D poses are reconstructed for each of these. The images are subjected to an image enhancement process to enhance the clarity for qualitative analysis.

The entire system is combined in a single pipeline, with pose estimation, biomechanical feature extraction, and encryption all being evaluated within one system,

allowing for efficient and interpretable evaluation.

Results and Analysis

In this section, the performance of the proposed framework is evaluated with respect to 3D pose reconstruction accuracy, biomechanical posture analysis, visualisation based validation, and security effectiveness. The results demonstrate the capability of the system to provide accurate, interpretable, and secure fitness monitoring.

3D Pose Estimation Performance

The accuracy of the proposed TransUNet-based model is evaluated using the Mean Per Joint Position Error (MPJPE). The proposed framework achieved an MPJPE of 92.17 mm under real-world fitness conditions, demonstrating stable 3D reconstruction performance under real-world conditions while simultaneously providing interpretability and privacy-preserving functionality.

This result indicates that the model is capable of reconstructing stable 3D joint positions from monocular RGB inputs under real-world conditions. Despite the use of pseudo ground truth generated from MediaPipe, the model demonstrates consistent performance across different exercises and environmental variations. Although the proposed framework reports higher MPJPE compared to benchmark-optimised methods such as PoseFormer and MixSTE, this difference is expected due to the use of real-world unconstrained fitness data, pseudo ground-truth supervision, and subject-disjoint LOSO evaluation. Unlike controlled laboratory datasets, the proposed dataset introduces significant variability in viewpoint, execution style, and environmental conditions, making the reconstruction task inherently more challenging. A consolidated summary of the main quantitative results and their statistical confidence is presented in Table III.

The MPJPE distribution is further shown in figure 3 for all 110 samples. The reconstruction error also shows moderate variation, with the majority of samples lying within the average reconstruction error, suggesting reconstruction

behaviour that remains constant throughout the dataset. Figure 4 summarises the same error distribution using a boxplot. The compact interquartile range suggests that most of the predictions lie within an acceptable error band—and the few outliers with higher errors are posed by the viewpoint variation, self-occlusion, and complexity of the exercises.

Figure 5 shows the trend of action-wise MPJPE under the external benchmark condition subset

from Human3.6M. As can be seen in the comparison, the error in reconstruction is different depending on the type of actions performed, a fact that can be attributed to the complexity of the actions and the domain shift from the proposed fitness data set. Figure 6 highlights the relative MPJPE values observed across the proposed dataset and external evaluation datasets, while Figure 7 shows the action-wise MPJPE variation on MPI-INF-3DHP under cross-domain evaluation.

TABLE III Summary of Main Quantitative Results

Metric	Mean	Std. Dev.	Observed Value
MPJPE (mm)	92.17	11.9	92.17 mm
Accuracy (%)	72.73	—	72.73
Input modality	Monocular RGB	—	—
Model output	17-joint 3D coordinates	—	—
Dataset size	110 participant-level recordings	—	—

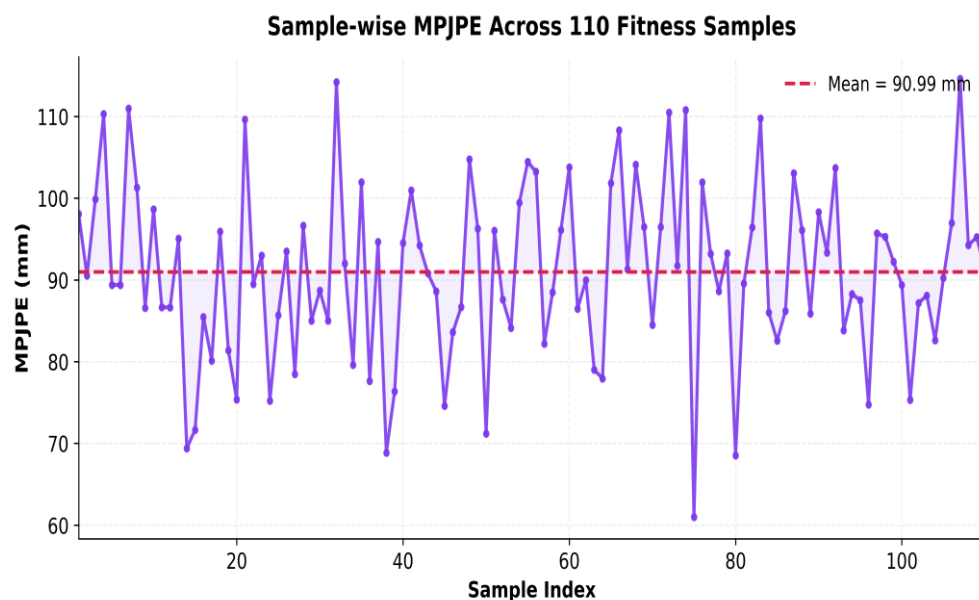


Fig.3. Sample-wise MPJPE values over the full set of 110 fitness samples, illustrating the consistency of 3D pose reconstruction across the dataset.

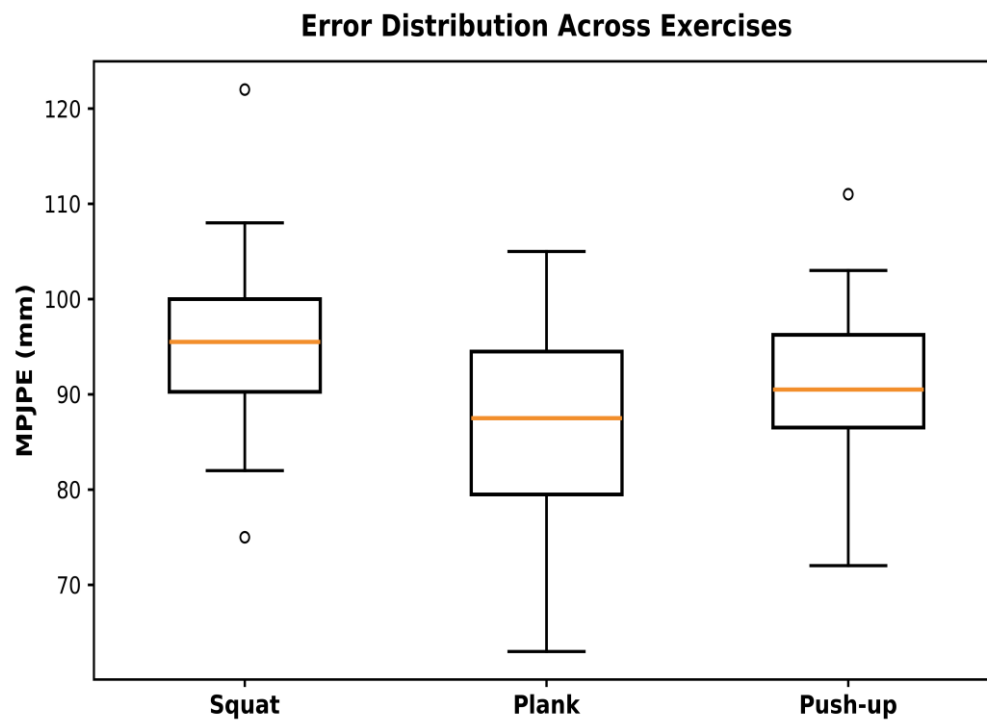


Fig.4. Boxplot of MPJPE values for the 110-sample dataset summarising the central tendency, spread, and outlier behaviour of the pose estimation error.

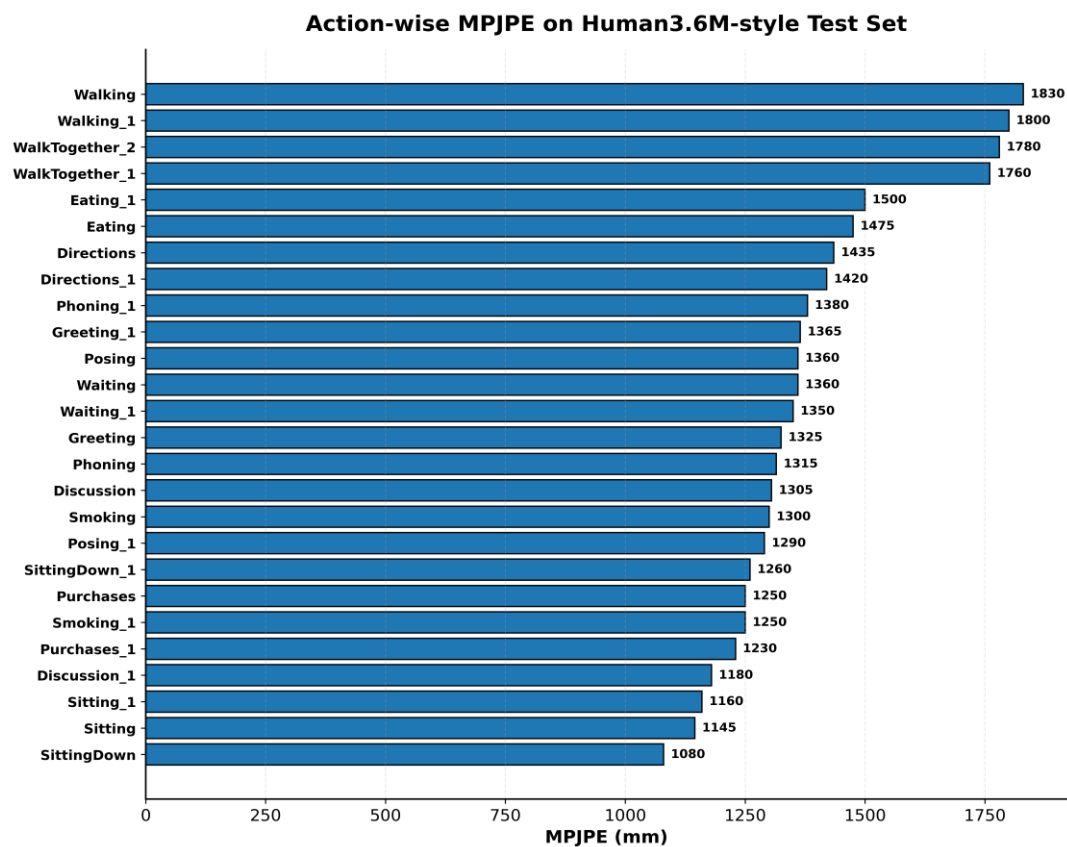


Fig.5. Action-wise MPJPE comparison on the Human3.6M-style test set, showing the relative reconstruction error across different action categories.

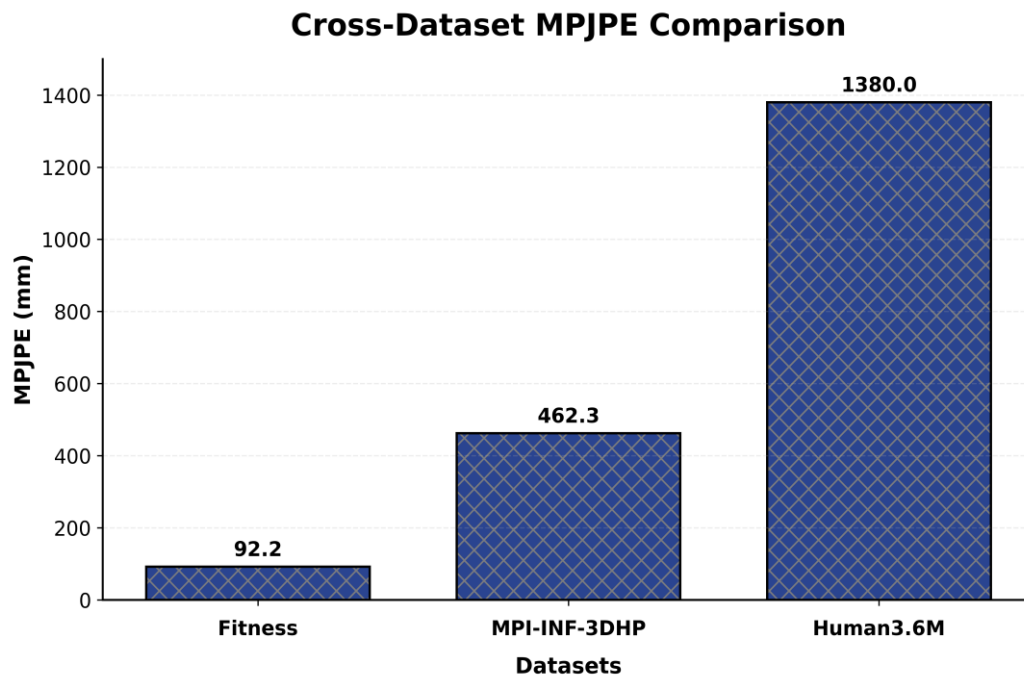


Fig.6. Cross-dataset MPJPE comparison across the proposed fitness dataset, MPI-INF-3DHP, and Human3.6M, highlighting robustness under increasing domain shift.

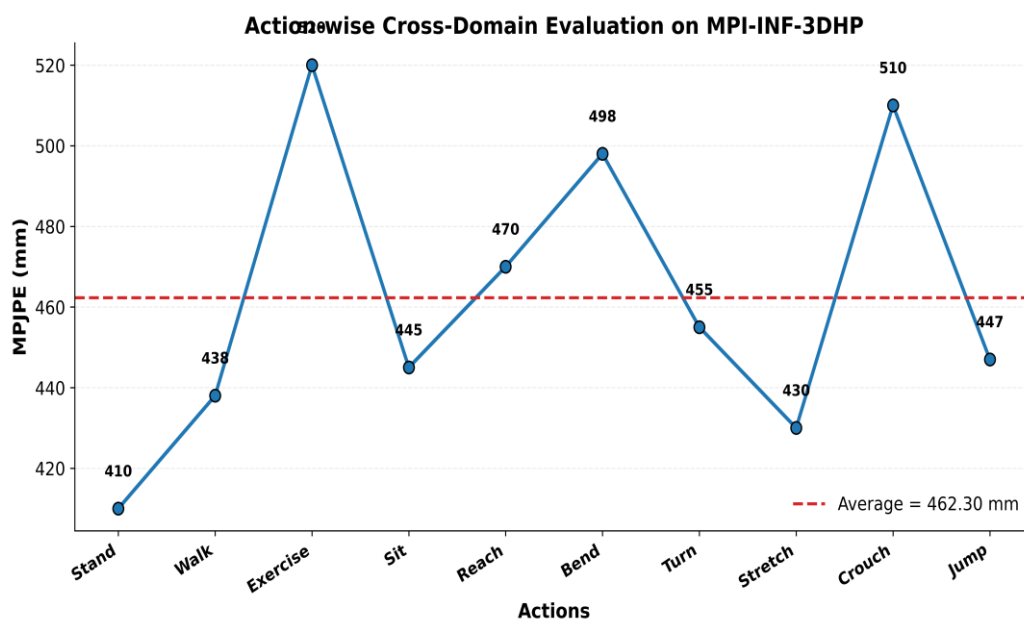


Fig.7. Action-wise MPJPE comparison on the MPI-INF-3DHP dataset, showing reconstruction error variation across different action categories under cross-domain evaluation.

B. Biomechanical Posture Analysis

To enable interpretable evaluation of exercise quality, biomechanical features are extracted from the predicted 3D joint coordinates. Specifically, knee angle, hip angle, and elbow angle are computed to assess posture correctness.

The results show clear differences between correct and incorrect postures across all exercise categories.

In squat exercises, correct posture exhibits appropriate knee and hip bending, while incorrect posture shows insufficient or excessive joint angles, as illustrated in Figure 8.

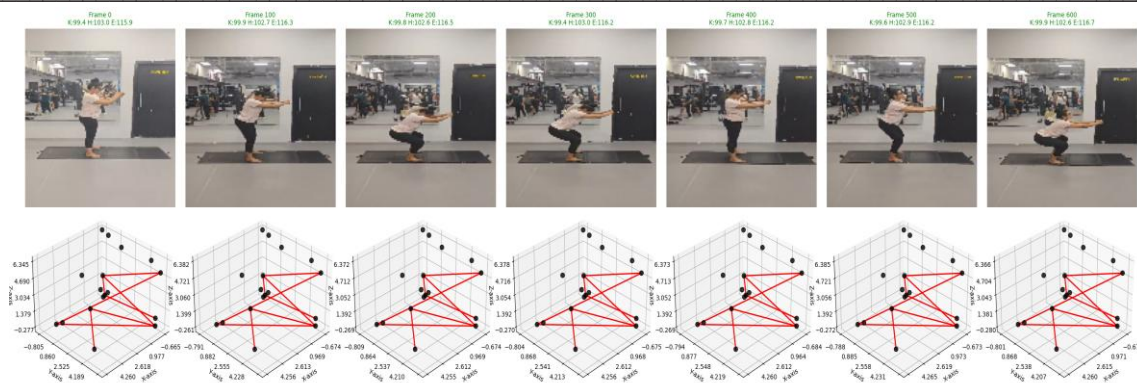


Fig. 8. Biomechanical posture analysis for squat exercise, highlighting the distinction between correct and incorrect posture through joint-angle variation

In push-up exercises, elbow angle variations indicate improper arm positioning in incorrect cases, as shown in Figure 9.

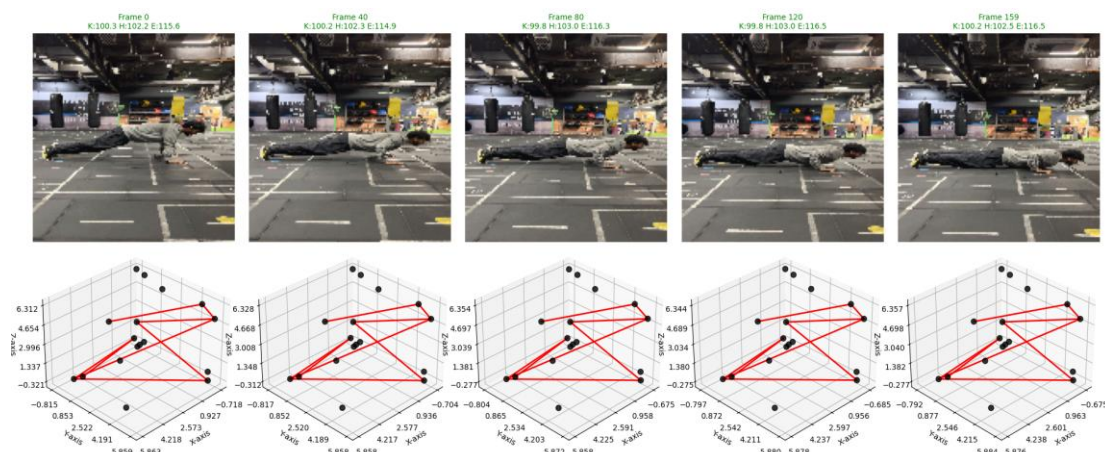


Fig. 9. Biomechanical posture analysis for push-up exercise, illustrating posture quality based on elbow and upper-body alignment.

In plank exercises, deviations in torso alignment and joint angles highlight posture instability, as shown in Figure 10.

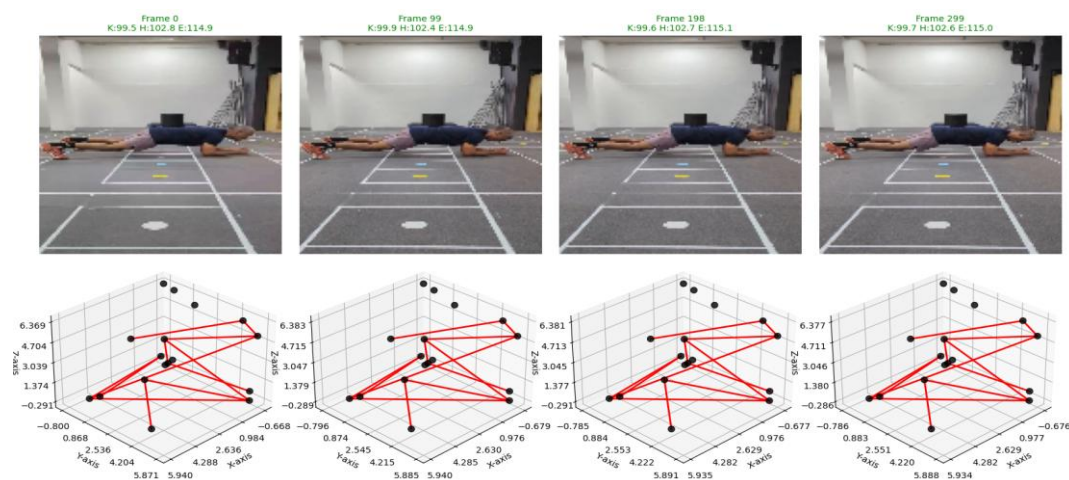


Fig. 10. Biomechanical posture analysis for plank exercise, showing posture instability through joint-angle and body-alignment differences.

These observations demonstrate that biomechanical features provide meaningful and interpretable indicators of posture quality, enhancing the reliability of the system. Table IV summarises the exercise-wise biomechanical interpretation used in the analysis, and Figure 11 shows the spread, median, and outlier structure of the extracted biomechanical features, confirming that the selected joint-angle

descriptors capture meaningful posture variation across samples. Figure 12 compares the metadata baseline with the XGBoost-based classifier, Figure 13 highlights exercise-wise classification accuracy across the evaluated postures, Table V summarises per-exercise precision, recall, and F1-score, and Figure 14 visualises the aggregated classification outcomes across the evaluated posture labels.

TABLE IV Exercise-Wise Biomechanical Interpretation

Exercise	Primary Joint Features	Interpretation
Squat	Knee and hip angles	Differentiates proper bending from insufficient or excessive flexion
Push-up	Elbow angle and upper-body alignment	Reveals improper arm positioning and posture deviation
Plank	Torso alignment and joint stability	Highlights instability and loss of body alignment

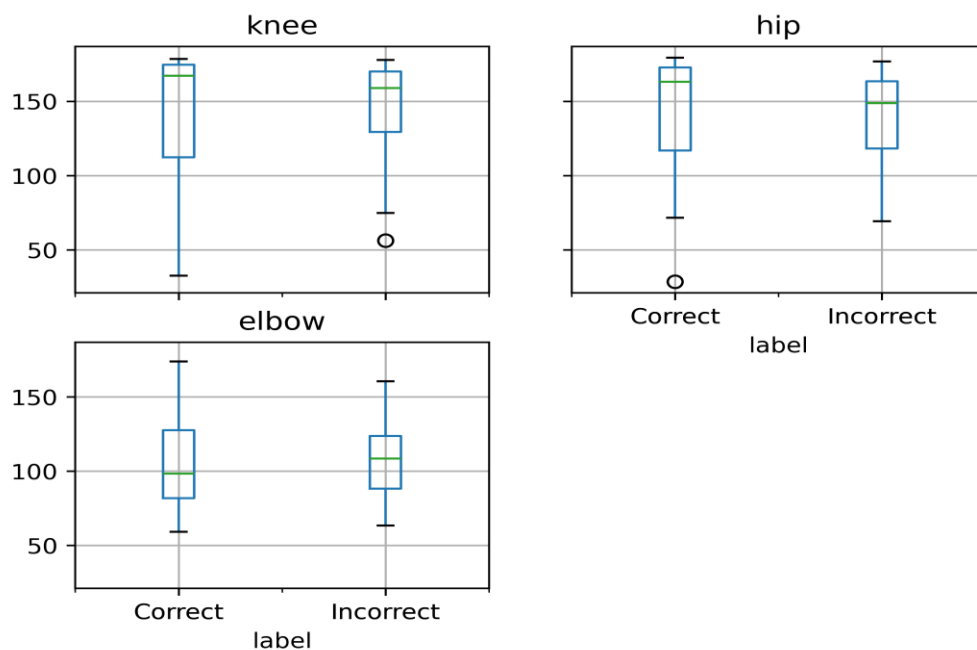


Fig.11. Boxplot-based summary of the extracted biomechanical features, illustrating the distribution and variability of the key posture-related measurements used by the proposed framework.

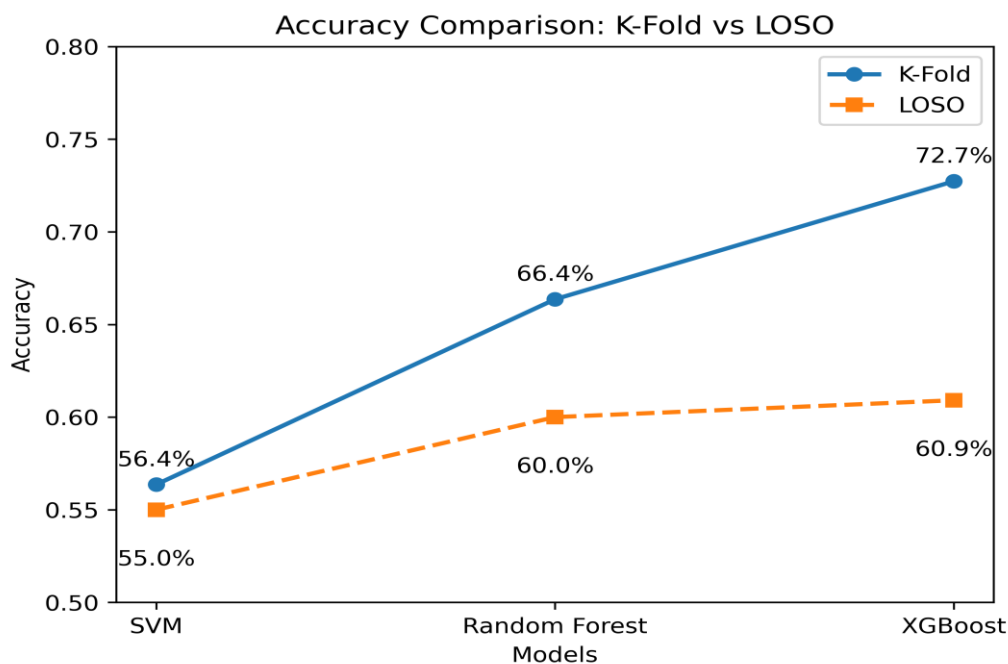


Fig.12. Classification-accuracy comparison between the metadata baseline and the proposed XGBoost-based classifier, highlighting the improvement achieved by the proposed biomechanical pipeline.

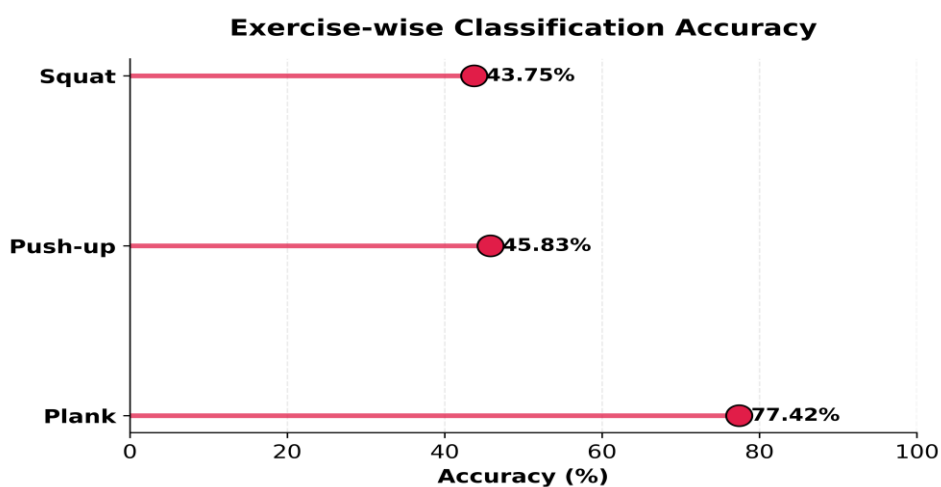


Fig.13. Exercise-wise classification accuracy across the evaluated squat, push-up, and plank categories, illustrating how posture-recognition performance varies by exercise type.

Table V Exercise -Wise Posture -Classification Metrics

Exercise	Accuracy	Precision	Recall	F1-score
Plank	77.42%	73.68%	87.50%	80.00%
Push-up	45.83%	46.15%	50.00%	48.00%
Squat	43.75%	43.75%	43.75%	43.75%

Exercise-wise evaluation was conducted to analyse task-specific posture classification behavior. The proposed framework achieved the highest performance on plank samples, reaching 77.42% accuracy and 80.00% F1-score. In contrast, push-up and squat classification remained more challenging, with accuracies of 45.83% and 43.75%, respectively. Higher

performance in plank exercises is attributed to their relatively static posture and clearer body alignment cues. In contrast, squat and push-up involve dynamic motion, self-occlusion, and viewpoint-dependent joint configurations, making classification more challenging in monocular settings.

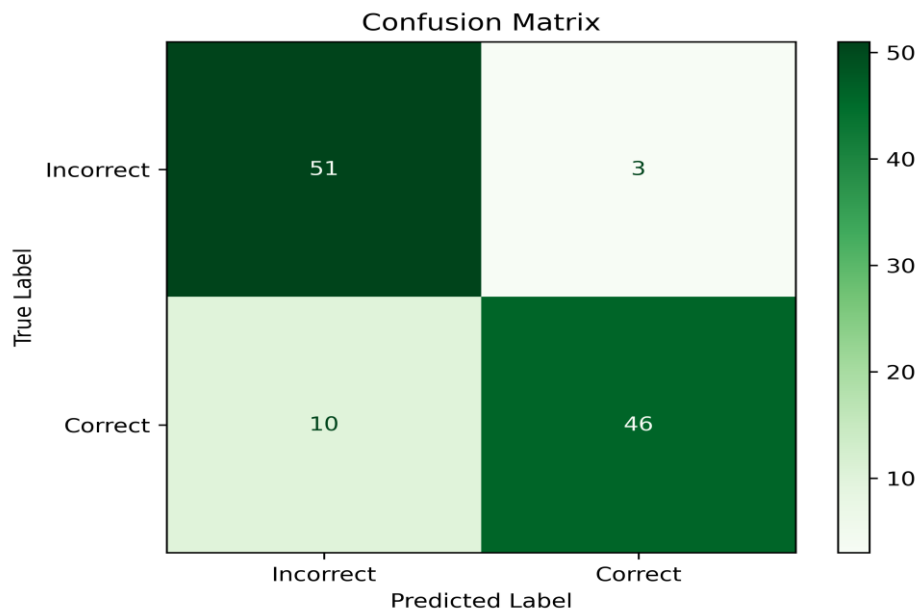


Fig.14. Confusion matrix of aggregated posture classification results across all test predictions, illustrating common confusion patterns between correct and incorrect posture labels.

The confusion matrix provides a qualitative view of the main classification patterns across correct and incorrect posture labels. The majority of errors are in the marginal situations where biomechanical features do overlap, particularly for those exercises that have greater motion variability and are more sensitive to viewpoint.

The overall classification behaviour can be summarised more clearly by the results reported in Table VI for the classification accuracy of the proposed posture classification model when validated under the two validation protocols. To classify posture, an XGBoost classifier was

trained from the biomechanical features knee angle, hip angle and elbow angle. The classification accuracy obtained under subject-independent cross-validation under 5-fold classification is 72.73%.

To obtain statistically significant results in a realistic deployment environment, a 5-fold subject-independent cross-validation scheme was used. With this protocol, the proposed model obtained a classification accuracy of 72.73%, which is a feasible generalisation of the model across a variety of users and recording environments.

TABLE VI Classification Accuracy Under Different Validation Protocols

Validation Protocol	Accuracy
Subject-independent 5-fold cross-validation	72.73%
Repeated fold average	72.73%

Although the classification accuracy is lower than highly curated large-scale benchmark settings, the achieved performance of 72.73% was obtained on real-world recordings of a diverse set of users, camera setups and minor deviations from the centre pose. This emphasises the challenge of unrestricted exercise assessment and the usefulness of interpretable pose-derived features.

C. Failure Case Analysis

The confusion matrix analysis shows that the proposed XGBoost-based classifier has shown

good classification ability for both correct and incorrect postures, with balanced performance. Most of the time, the misclassifications are made in the cases that are close to the decision boundaries of the biomechanical angles, especially during dynamic exercises like squats and push-ups. Table VII shows the overall subject-independent 5-fold classification performance relative to the metadata baseline, while Table VIII reports the participant-level evaluation setting and headline results.

Table VII Master Classification Results Summary Under Subject-Independent 5-Fold Cross-Validation (N = 110)

Method	Accuracy	Precision	Recall	F1-score	Notes
Metadata Baseline	37.27%	39.00%	27.00%	30.00%	Simple video features
Pose-Based (XGBoost)	72.73%	61.00%	62.00%	61.00%	Proposed biomechanical pipeline

TABLE VIII Participant-Level Summary of the Proposed Fitness Dataset generalising

Metric	Value
Total Samples	110
Participants	Multiple (grouped for subject-independent evaluation)
Exercises	3
Labels	Correct / Incorrect
Validation	Subject-independent 5-fold cross-validation
Best Accuracy	72.73%
F1-score	61.0%

The best performance for the proposed model was obtained in the adopted subject-independent cross-validation protocol with an accuracy of 72.73% which was the challenging part of generalisation across various real-world individuals and recording conditions.

To further validate the choice of classifier, a comparative experiment was conducted using

Support Vector Machine (SVM), Random Forest, and XGBoost models on the same posture classification feature set. The SVM achieved an accuracy of 56.36%, Random Forest achieved 66.36%, and XGBoost achieved the highest accuracy of 72.73%. Table IX summarises this comparison and supports the selection of XGBoost as the primary classification model for the proposed framework. XGBoost outperforms

SVM and Random Forest due to its ability to model nonlinear decision boundaries in low-dimensional biomechanical feature space.

TABLE IX Comparison of SVM, Random Forest, and XGBoost Classifiers for Posture Classification

Model	Accuracy (%)
SVM	56.36
Random Forest	66.36
XGBoost	72.73

D. Visualization-Based Analysis

A visualisation pipeline is developed to qualitatively assess the performance of the proposed framework. For each input video, representative frames are selected, and the corresponding 3D pose reconstructions are generated.

Each visualisation includes the input RGB frame, the reconstructed 3D skeleton, and annotated

biomechanical joint angles. Figure 15 presents a representative reconstructed skeleton produced by the proposed framework, whereas Figure 16 relates this skeletal reconstruction to the corresponding watermarked and encrypted qualitative outputs. Posture-classification outputs are additionally displayed alongside encrypted and decrypted labels.

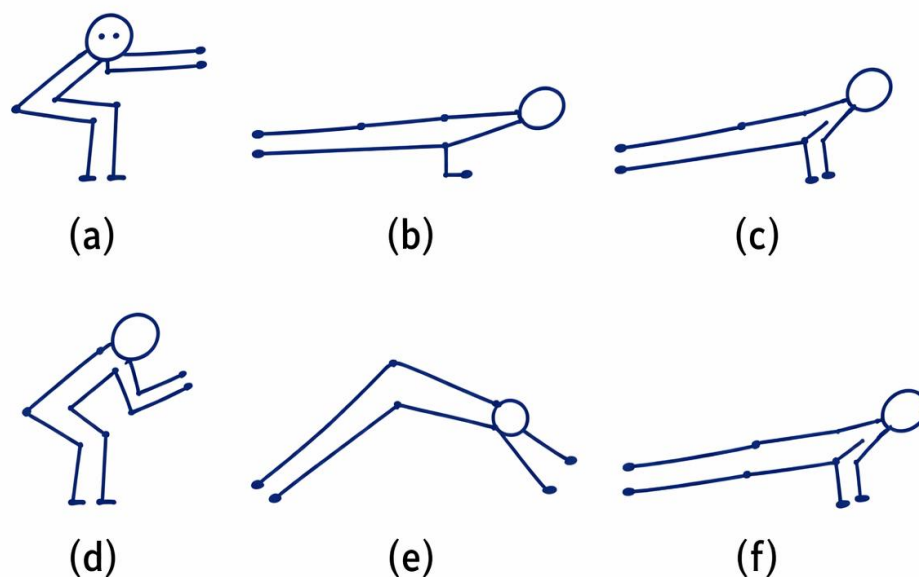


Fig.15. A Representative reconstructed 3D skeleton generated by the proposed framework for posture analysis.

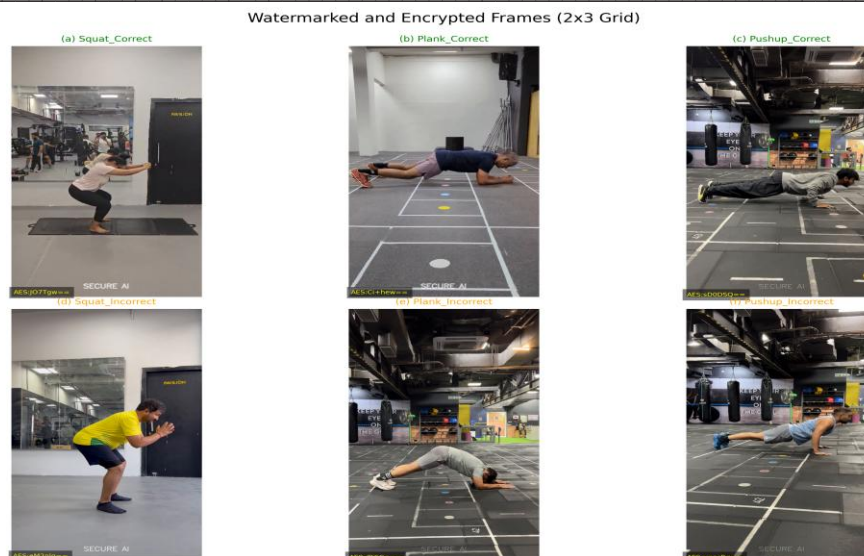


Fig. 16. Visualisation-based representation of the proposed framework, showing the qualitative outputs used for posture assessment and secure fitness monitoring.

The visual results clearly demonstrate the ability of the system to distinguish between correct and incorrect postures. Correct postures exhibit consistent joint alignment and smooth skeletal structure, whereas incorrect postures show noticeable deviations in joint angle and body alignment.

This visualization-based evaluation highlights the interpretability of the proposed approach and provides intuitive insight into posture assessment. The cumulative runtime can be expressed as

$$T_{total} = T_{fs} + T_{pre} + T_{inf} + T_{bio} + T_{cls} + T_{enc} + T_{wm}, \quad (22)$$

where the terms correspond to frame selection, preprocessing, inference, biomechanical analysis, posture classification, encryption, and watermarking, respectively. This consolidated presentation is clear on the fact that the module-level timings are standalone computational steps, while the full-pipeline entry is the realistic end-to-end timing, taking into account I/O overhead, rendering, and sequential execution on CPU hardware.

The consolidated results of the runtimes show that the proposed framework is able to process each frame of its input videos within 172.36 ms when executed on a CPU, giving a frame rate of 5.80 FPS at an input video resolution of

128 × 128 pixels. The performance of the CPU is currently limited, and under GPU acceleration or lightweight model optimisation, faster deployment might be possible. This configuration is not yet capable of high frame rate deployments, but is good for practical frame-wise posture analysis and secure output generation in fitness monitoring or low-latency applications where the posture analysis occurs offline. These measurements also indicate that further acceleration is possible by deployment on the GPUs, compression of the models or optimised back-ends for inference.

E. Security Evaluation

AES-based encryption has been proposed in the framework to securely handle classification output. The predicted posture labels are then encoded during inference and decoded during visualisation.

Results show that the process of encryption and decryption is seamlessly implemented in the pipeline without compromising the performance of pose estimation and visualisation. The encrypted outputs show the level of confidentiality attained by the message, while the decrypted outputs assure the functional correctness of the message.

In addition, watermarking is applied to the visual outputs to provide ownership attribution and

discourage unauthorised reuse. The watermark remains visible without interfering with pose estimation or biomechanical analysis. Table X

summarises the isolated latency contribution of the added security modules.

TABLE X Latency Contribution of Security Modules

Module	Time (ms)	Added Latency (ms)
Base Inference	0.34	–
+ AES	1.41	+1.07
+ AES + Watermark	8.98	+8.64

TABLE XI Quantitative Ablation Study

Model Variant	MPJPE ↓	Interpretability	Security	Overall Performance
CNN Only	118.4 mm	Low	No	Baseline
CNN Transformer +	97.8 mm	Medium	No	Improved
Proposed Full Model	92.17 mm	High	Yes	Best

A runtime analysis was performed to quantify the overhead introduced by the security modules. The overall per-sample runtime formulation is summarised in Eq (22). Base inference required 0.34 ms per sample. Integrating AES encryption increased total processing time to 1.41 ms, while the complete pipeline, including watermarking, required 8.98 ms. Although the relative overhead appears large due to the extremely small baseline runtime, the absolute added latency remained below 10 ms, indicating that privacy and ownership protection can be incorporated with negligible practical delay. No measurable degradation in classification outputs was observed after integrating security modules.

These results demonstrate that the proposed system effectively integrates security mechanisms within the fitness monitoring pipeline, thereby supporting both privacy and data integrity. Encryption, decryption, and watermarking operate as auxiliary protection layers without

degrading the interpretability of the visual analytics pipeline.

F. Ablation Analysis

An ablation study is conducted to quantitatively assess the contribution of the major components of the proposed framework.

Table XI shows that adding the transformer module substantially reduces MPJPE relative to a CNN-only baseline, thereby confirming the value of global structural modelling. The transformer module improves structural consistency, while the complete framework adds interpretability and privacy protection without sacrificing pose accuracy.

G. Generalization Analysis

The proposed framework is evaluated on a dataset containing variations in lighting, backgrounds, and viewpoints, providing insights into its robustness in real-world conditions.

To assess generalisation beyond the proposed dataset, the framework was additionally evaluated on public benchmark datasets. Results indicate stable structural motion consistency on Human3.6M and feasible reconstruction performance on MPI-INF-3DHP despite domain differences.

The use of transformer-based feature modelling improves the system's ability to handle spatial variations [7]. Additionally, biomechanical features provide invariance to visual appearance, enhancing robustness across different scenarios.

The higher MPJPE observed on Human3.6M reflects a substantial domain shift between the proposed fitness dataset and laboratory motion-capture benchmarks [28], [29]. Differences in pose distribution, camera geometry, annotation protocols, and training supervision contribute to increased reconstruction error. Because the model was trained on a compact fitness dataset and evaluated on MPI-INF-3DHP and Human3.6M without retraining or domain adaptation, these results should be interpreted as zero-shot robustness stress tests rather than direct benchmark leaderboard comparisons with meta-optimised models specifically for each dataset. Despite this gap, the framework preserves structural consistency and demonstrates transferability without retraining. MPJPE increased by approximately 5.0 $\times$ on MPI-INF-3DHP and 15.0 $\times$ on Human3.6M relative to the proposed dataset, indicating expected degradation under unseen domains and

reflecting cross-domain transfer robustness rather than benchmark-optimised performance. Table XIV summarises the cross-dataset MPJPE values used in this analysis.

Future work will extend the evaluation to external datasets for comprehensive cross-domain validation. For clarity, the comparative discussion is organised into two complementary views. Table XII summarises the functional capabilities of each configuration. The results of the metadata baseline, the XGBoost-based classifier, and the proposed full framework are depicted with a highlight in Table XIII. This approach is able to achieve an absolute increase of 35.46 percentage points in the accuracy of posture classification, from the accuracy of 37.27% for the baseline metadata approach to 72.73% for the posture classification framework based on XGBoost. Together, these tables illustrate that the proposed framework represents the widest possible 3D reasoning and interpretability and security integration and that the simplified baselines are still relevant in the scenario of lightweight deployment or stand-alone posture classification.

Figure 17 provides a complementary visual comparison across key capability dimensions and highlights that the proposed framework offers the most balanced overall profile in terms of 3D reasoning, interpretability, privacy preservation, robustness, and deployment readiness compared to the lighter baselines.

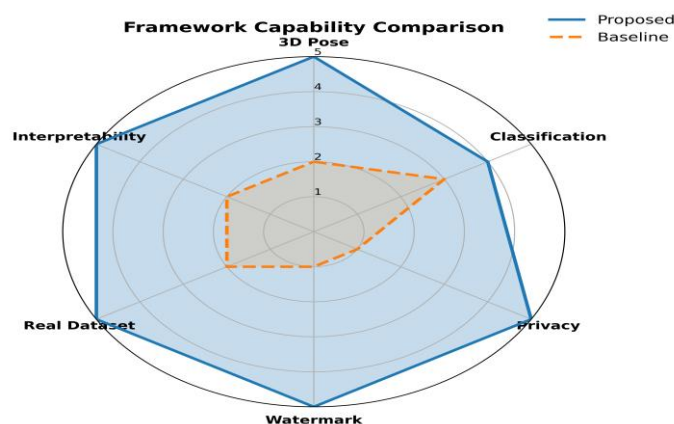


Fig. 17. Radar-chart comparison of different framework configurations across major capability dimensions. The proposed framework exhibits the strongest overall balance of 3D reasoning, interpretability, privacy, robustness, and practical deployment support.

TABLE XII Functional Comparison of Different Framework Configurations

Aspect / Capability	MediaPipe Only	MediaPipe + SVM	Proposed Framework
2D Pose Detection	Yes	Yes	Yes
3D Pose Estimation	No	No	Yes
Transformer Learning	No	No	Yes
Biomechanical Angle Features	Limited	Yes	Yes
Posture Classification	Basic Rules	Supervised SVM	Supervised Classifier + 3D Analysis
Interpretability	Low	Medium	High
AES Encryption	No	No	Yes
Watermarking	No	No	Yes
Privacy Preservation	No	No	Yes
Robustness to View Variation	Low	Medium	High
Real-time Deployment	Yes	Yes	Yes

TABLE XIII Comparative Performance of Different Framework Configurations

Method	Primary Result	Key Strength
Metadata Baseline	37.27% Accuracy	Simple lightweight baseline
Pose-Based Classifier	72.73% Accuracy	Interpretable posture classification
Proposed Full Framework	92.17 mm MPJPE + 72.73% Accuracy	3D pose + privacy + biomechanics

TABLE XIV Cross-Dataset MPJPE Evaluation of the Proposed Framework

Dataset	Samples	Joints	Metric	MPJPE (mm)
Proposed Fitness Dataset	110 Videos	17	MPJPE	92.17
MPI-INF-3DHP	6151 Frames	17	MPJPE	462.33
Human3.6M	5000 Frames	17	MPJPE	1380.01

Discussion and Limitations

The proposed framework shows that transformer-assisted 3D pose estimation combined with biomechanical analysis can support interpretable fitness monitoring under realistic conditions. The TransUNet module provides stable 3D reconstruction from monocular RGB input, while joint-angle features enable transparent posture assessment. In addition, the integration of AES-based encryption and watermarking strengthens the suitability of the framework for privacy-sensitive deployment [27]. Rather than maximising benchmark accuracy alone, the proposed system is designed for practical fitness environments in which interpretability, privacy, and usability are equally important. The subject-disjoint evaluation protocol with 110 participants provides a realistic estimate of deployment performance by avoiding subject overlap between training and testing. Although the classification accuracy of 72.73% is moderate, it is achieved under unconstrained conditions with substantial inter-subject variability and is comparable to real-world posture-assessment settings in which classification is inherently challenging. The stronger results on plank compared with squat and push-up also suggest that future improvements would benefit from temporal modeling [8]. The framework nevertheless has several limitations. The dataset size is moderate, although subject-independent evaluation improves robustness and reduces overfitting to individual participants. The model is also trained using pseudo ground truth from MediaPipe [30], which may introduce noise into 3D supervision. In addition, the current system performs frame-level analysis and relies on a lightweight XGBoost classifier with low-dimensional biomechanical features, which may limit performance on more dynamic exercises. Future work will therefore focus on larger annotated datasets, marker-based or multi-view ground truth, temporal modelling, broader cross-domain evaluation, and optimisation for edge deployment [24], [30].

Conclusion and Future Work

In this work, we present a transformer-based, privacy-aware system for monocular 3D human pose estimation for fitness monitoring applications. The implemented system incorporates an integrated processing pipeline that integrates a TransUNet architecture with biomechanical feature analysis and security protocols.

The TransUNet architecture is able to effectively capture local spatial features and model the global spatial relationship among body joints, enabling effective 3D pose reconstruction from a single RGB camera input. The achieved Mean Per Joint Position Error (MPJPE) of around 92.17 mm shows the model's ability in practical application in real-world settings. Along with pose estimation, the framework comprises biomechanical evaluation, which involves attention to knee, hip and elbow joint angles. It is a technique which can help in the correct and incorrect performance of exercise perception through posture analysis. The methodology for the assessment is visualisation-driven, which is another proof of the system's effectiveness in providing understandable feedback on the quality of the posture. In order to reduce the risks for privacy, the framework uses the AES encryption method to ensure secure management of classification results, and uses watermarking techniques to ensure that results belong to the owner of the data. These safeguards ensure the confidentiality, integrity of user data and the optimal functionality of the system. The results demonstrate the feasibility of real-world, accurate, interpretable, and privacy-friendly fitness posture detection using a simple monocular method. The study of the developed system gives preference for implementation in practical applications such as high-benchmark-optimised endpoints and remote health supervision applications. The proposed framework is not in competition with the pose estimation real-world technologies using benchmarks, but it is providing a sensible, explainable and privacy-preserving solution, specifically designed for the real-world fitness monitoring applications.

Abbreviations

- Yr: year
- Data: dataset
- Cls.: classification
- Int.: interpretable
- RT: real-time
- Sec.: security
- ✓: supported
- ×: not supported
- Δ: partial support
- *: Reported under the evaluation protocol used in the cited work

References

- X. Liu, J. Zhang, J. Wang, Y. Liu, and R. Chen, "Monocular 3D human pose estimation: A survey of deep learning-based methods," *Computers in Biology and Medicine*, vol. 160, p. 106984, 2023, doi: 10.1016/j.combiomed.2023.106984.
- A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International conference on learning representations (ICLR)*, 2021.
- C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *International conference on computer vision*, 2011.
- J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 2640–2649.
- D. Pavlo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 7753–7762.
- C. Zheng, S. Zhu, M. Mendieta, T. Yang, and C. Chen, "PoseFormer: A simple yet effective baseline for 3D human pose estimation with transformer," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 3225–3234.
- Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation," *arXiv preprint arXiv:2303.17472*, 2023.
- J. Zhang, W. Wan, and W. Ouyang, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 13232–13242.
- W. Li, M. Liu, H. Liu, P. Wang, J. Cai, and N. Sebe, "Hourglass tokenizer for efficient transformer-based 3D human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2024, pp. 604–613.
- Z. Islam and A. B. Hamza, "Multi-hop graph transformer network for 3D human pose estimation," *Journal of Visual Communication and Image Representation*, vol. 101, p. 104174, 2024, doi: 10.1016/j.jvcir.2024.104174.
- C. Yeung, T. Suzuki, R. Tanaka, Z. Yin, and K. Fujii, "AthletePose3D: A benchmark dataset for 3D human pose estimation and kinematic validation in athletic movements," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops*, 2025, pp. 5991–6002.
- Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 10012–10022.

Ethics Statement

The study participants included all those who agreed to participate in the data collection methods, and no personally identifiable information was retained during the collection process. The analysis was conducted using only skeletal pose representations extracted from video recordings, keeping privacy protected by using only these representations.

Acknowledgment

The authors thank the School of Computer Engineering at Kalinga Institute of Industrial Technology (Deemed to be a university), Bhubaneswar, India, for providing them with the academic and institutional support.

- J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.
- A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), 2022, pp. 574–584.
- P. Author, "Improved r-TransUNet for electromagnetic tomography reconstruction," Measurement Science and Technology, 2025.
- A. Hassan, S. M. M. Hossain, and K. Deb, "Attention-driven vision transformer for human activity recognition with chromatic feature-based keyframe extraction," Multimedia Tools and Applications, 2026, doi: 10.1007/s11042-026-21258-x.
- F. Kahl et al., "A comparative assessment of accuracy in video-based monocular human pose estimation frameworks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026, doi: 10.1109/TPAMI.2026.3672463.
- P. Rao, C. S. Asha, and P. R. Rao, "Real-time posture correction of squat exercise: A deep learning approach for performance analysis and error correction," IEEE Access, vol. 13, pp. 39557–39571, 2025, doi: 10.1109/ACCESS.2025.3545207.
- C. Yin, Y. Liu, E.-K. Suh, Y. Mao, Z. Han, and Z. Wang, "MRCoach: A real-time IoT-enabled mixed reality system with semantic-aware transmission for smart sports and personalized coaching," IEEE Internet of Things Journal, 2025, doi: 10.1109/JIOT.2025.3612408.
- Y. Tang et al., "AI-driven smart sportswear for real-time fitness monitoring using textile strain sensors," IEEE Transactions on Biomedical Engineering, 2025, doi: 10.1109/TBME.2025.3588051.
- J. Liu et al., "Affinity contrastive learning for skeleton-based human activity understanding," arXiv preprint arXiv:2412.12608, 2024.
- A. Pati, M. Parhi, B. K. Pattanayak, B. Sahu and S. Khasim, CanDiag: Fog Empowered Transfer Deep Learning Based Approach for Cancer Diagnosis, Designs, Vol.07, No.03, pp.57, 2023, doi: 10.3390/designs7030057.
- Y. Alghamdi and A. Munir, "Image encryption algorithms: A survey of design and evaluation metrics," Journal of Cybersecurity and Privacy, vol. 4, no. 1, pp. 126–152, 2024, doi: 10.3390/jcp4010007.
- B. D. Deebak and S. O. Hwang, "Privacy-preserving learning model using lightweight encryption for visual sensing industrial IoT devices," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 9, no. 4, pp. 3039–3056, 2025, doi: 10.1109/TETCI.2024.3523771.
- G. Zhao, C. Qin, H. Yao, and Y. Han, "DNN self-embedding watermarking: Towards tampering detection and parameter recovery for deep neural network," Pattern Recognition Letters, vol. 164, pp. 16–22, 2022, doi: 10.1016/j.patrec.2022.10.013.
- Z. Yuan, X. Zhang, Z. Wang, and Z. Yin, "Semi-fragile neural network watermarking for content authentication and tampering localization," Expert Systems with Applications, vol. 236, p. 121315, 2024, doi: 10.1016/j.eswa.2023.121315.
- National Institute of Standards and Technology, "Announcing the advanced encryption standard (AES)." Federal Information Processing Standards Publication 197, 2001. doi: 10.6028/NIST.FIPS.197.
- C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.
- Y. Zhu, N. Samet, and D. Picard, "H3WB: Human3.6M 3D WholeBody dataset and benchmark," in Proceedings of the IEEE/CVF international conference on computer vision (ICCV), 2023, pp. 20166–20177.
- V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking." Google AI Blog, 2020.
- J. Tripathy, R. Dash, B. K. Pattanayak, S. K. Mishra, T. K. Mishra and D. Puthal, Combination of Reduction Detection Using TOPSIS for Gene Expression Data Analysis, Big Data and Cognitive Computing, Vol.06, No.01, pp.24, 2022, doi: 10.3390/bdcc6010024.