IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 August 2023, accepted 12 September 2023, date of publication 18 September 2023, date of current version 4 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316009

RESEARCH ARTICLE

AI Trainer: Autoencoder Based Approach for Squat Analysis and Correction

MUKUNDAN CHARIAR, SHREYAS RAO, ARYAN IRANI, SHILPA SURESH, (Senior Member, IEEE), AND C S ASHA, (Member, IEEE)
Department of Mechatronics, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India
Corresponding author: C S Asha (asha.cs@manipal.edu)

ABSTRACT Artificial intelligence and computer vision have widespread applications in workout analysis. It has been extensively used in sports and the athlete industry to identify errors and improve performance. Furthermore, these methods prevent injuries caused by a lack of instructors or costly infrastructure. One such exercise is the squat, which is a movement in which a standing person descends to a posture with their torso vertical and their knees firmly bent, then returns to their original upright position. Each person's squat is distinct, with varying limb lengths causing their form to change when observed. It has been observed that the mobility of various joints and muscular strength have a role in this. A squat improves the user by increasing overall leg strength, strengthening knee and hip joints, and lowering the risk of heart disease due to cardiovascular development. This paper presents a method for classifying squat types and recommending the right squat version. This study uses MediaPipe and a deep learning-based technique to decide if squatting is good or bad. A stacked Bidirectional Gated Recurrent Unit (Bi-GRU) model with an attention layer is proposed to consistently and fairly assess each user, categorizing squats into seven classes. This stacked Bi-GRU model with an attention unit is then compared to other cutting-edge models, both with and without the attention layer. The model outperforms other models by attaining an accuracy of 94% and is demonstrated to work the best and most consistently for our dataset. Furthermore, the individual executing the incorrect squat is corrected to the best of their ability, depending on their performance and body proportions, by providing the correct form.

INDEX TERMS Action quality assessment, attention, computer vision, curve fitting, gated recurrent unit, pose estimation, squat.

I. INTRODUCTION

An incorrect type of exercise might result in injury; thus, the exercise should be performed under professional supervision. A squat is a standing exercise in which a person descends to a posture with a vertical torso and fully bent knees, then restores to a normal standing position. Squats are one of the most challenging exercises. Each individual will have a unique squat; when observed, the lengths of their limbs will cause their form to shift. It has been observed that the flexibility of different joints and the power of respective muscles influence the type of squat performed. A person with a long femur but a short torso could not squat as deeply as someone with a short femur but a long torso. This does not imply that a person's squat is unacceptable; rather, it is a variant of a proper squat. Similarly, a person with an anthropometry such that their hip opens wider than usual would be able to squat to depth with their feet comfortably positioned apart. Still, a normal individual would be unable to do so.

The squat has several advantages, including increased overall leg strength, stronger knee and hip joints, and a decreased risk of heart disease because of cardiovascular development. Squatting regularly strengthens the legs, muscles and bones, lowering the risk of osteoporosis. A plethora of literature deals with models using sensors to provide

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti.

exercise assistance [1]. Sensors like accelerometers record the movement or vibration of each joint to classify the activity. In addition, a lot of literature addresses sensors, including cameras and Kinect, to accurately retrieve the joint positions of the user [1]. As a result, posture estimation, implemented in multiple studies, is a crucial part of workout evaluation. AI Fitness Trainer using MediaPipe [2] utilizes MediaPipe as a pose estimation algorithm and uses hard-coded methods to evaluate squats. The metric is the angle formed at the knee, between the hip and the ankle. Consequently, it only accepts videos taken from the side and classifies squats as good or bad. Different beginner and advanced modes are also provided, changing the threshold to classify the squats. SquatDepth [3] utilizes the MediaPipe in the same way as in [2] and has hard-coded its classification method. The metric uses the y coordinate of the hip and the y coordinate of the knee. However, it also accepts videos taken from the side. Feedback is given visually via three dots that turn green if the squat is classified as good. Otherwise, the dots remain red. Squatevaluation [4] uses a method similar to [5] to perform pose estimation. Authors further use a 3 Dimensional (3D) Convolutional Neural Network (CNN) to classify squats. This method can only classify squats into two types, good or bad, and tends to classify bad squats as good squats. This method accepts videos taken from the front and the side for classification. Squat-Classification-And-Counting [6] uses the pose estimation method used in [7] and a custom-defined Neural Networks (NN) regressor to classify squats. The squats are classified according to depth and have three classes: quarter squat, half squat, and full squat. This method, again, can only accept videos taken from the side and fails to classify squats otherwise. Standard-Squat-Posture-Classifier [8] uses MediaPipe for pose estimation, and Support Vector Machine (SVM) classifier is used to classify squats. The squats are classified into six classes, and the classification takes place in real-time. This allows for real-time feedback, which is displayed to the user. Videos taken from the side and the front are accepted by this method. IVU [9] uses MediaPipe to extract pose and an Long Short-Term Memory (LSTM) to classify squats. Squats are classified into seven classes based on [10]. This method also classifies squats based on videos taken from the front and side.

The following observations were made based on the previous works. The squat analysis problem combines several techniques in Human Action Recognition (HAR), Pose Estimation, Action Imitation, and Classification.

1) **Human Action Recognition:** HAR is recognizing actions using computing methods. Multiple methods, such as Hidden Markov Model (HMM)s, Bayesian learning, NNs, SVM, etc., have been used to perform HAR.
2) **Pose Estimation:** Pose estimation is utilized to gather data on joint angles and limb lengths. Multiple algorithms such as OpenPose, MediaPipe Pose, BlaisePose, You Only Look Once v7 (YOLOv7) pose, VoxelPose, etc., are used to provide the pose data in 3D coordinate format for multiple points on the body.
3) **Exercise Evaluation and Analysis:** Exercise evaluation using methods such as NN, HMM, SVM, etc., has been done previously. Algorithms are built to evaluate exercises after extracting features using Pose Estimation algorithms. The metrics vary for each exercise.
4) **Action Imitation:** Action Imitation is a relatively new field. Actions have been imitated by periodic repetition of the action and estimating the action by using curves to define the motion of a point on the body.

A. ACTION RECOGNITION

Chen et al. [11] provided a technique for identifying 10 different activities using an HMM and star skeletonization as a typical description of human posture. A real-time, low-cost technique to detect falls has been presented by Dubois and Charpillet [12], which employs HMM to identify seven out of eight possible actions successfully. Using a combination of the Kinect v2 and the star skeleton algorithm, Hai and Kha [13] utilized HMM to identify seven more actions and classify activities in both indoor and outdoor settings. By combining HMM with NN-based technologies, Singh et al. [14] presented an automated video surveillance model that can differentiate between suspicious and non-suspicious actions in a monitored environment. A human action recognition technique that makes use of angle and coordinate-based features in addition to a multivariate continuous Gaussian Mixture Model (GMM) classifier is proposed by Hachaj and Ogiela in [15]. A methodology for recognizing and predicting human movements in human-robot interactions has been developed [16]; it needs to be emphasized that this framework accurately predicts the near future; however, it cannot represent a whole motion. A Bayesian approach is presented by Madabhushi and Aggarwal [17] and uses a monocular grayscale image sequence to track the subject's head movement throughout a series of frames to identify human activity. A human activity identification system for mobile cameras is presented [18]. Human activity recognition follows the categorization of the body's posture by considering all information on the subject's pose, location, and time elapsed.

Wu et al. [19] describe a method called Deep Dynamic Neural Network (DDNN) for multi-modal gesture recognition. A 3D CNN model that can extract features of the spatial and temporal dimensions is proposed [20], which performs 3D convolutions. The model generates multiple information channels from the input frames and combines all of them for the final feature representation. Paulose et al., in [21], describe a method using a Recurrent Neural Network (RNN) to classify actions with star skeletons. An architecture employing CNN, LSTM and temporal-wise attention to determine human actions is presented in [22]. Putra et al. [23], propose 3 LSTM models and evaluate

them against 4 other LSTM models for HAR, evaluated on the Weissman Dataset. The MediaPipe is used as the preprocessing algorithm, which extracts 1662 points from the video. Sudhakar Yadav et al. [25] propose an LSTM-based Activity Recognition System to compare a nondescript pose estimation algorithm with MediaPipe's pose estimation algorithm. In addition, they compare RNN with the LSTM. The LSTM with MediaPipe performs the best. Majd and Safabakhsh [26] propose a correlational convolutional LSTM, or Correlational Convolutional Long Short Term Memory (C^2LSTM), which utilizes correlational and convolutional layers along with LSTM layers to provide a complete model for HAR. Zhang et al. in [27] compare Convolutional Long Short Term Memory (ConvLSTM)s and Fully Connected Long Short Term Memory (FC-LSTM)s with different attentions. They design and compare a Spatial-Temporal Dual Attention Network (STDAN) for HAR to existing models. Authors, in [28], propose a Spatio-Temporal Long Short Term Memory (ST-LSTM) with Trust Gates to perform 3D HAR. The ST-LSTM is a variation of the traditional LSTM network that considers the 3D position of each node on the body, extracted via Pose Estimation algorithms. Liu et al. [29] propose a Global Context-Aware Attention Long Short Term Memory (GCA-LSTM) network for 3D HAR. The GCA-LSTM is a variation of the traditional LSTM network, much like ST-LSTM, which takes into consideration that while recognizing actions, humans focus on some part of the body or, in other words, pay 'attention' to it. A human action scoring model that is unsupervised and autoencoder-based is presented in [30], which detects and understands the temporal patterns of the human pose across multiple frames by utilizing a sequence-to-sequence model. Doan [31] implement HAR using NNs on a Raspberry Pi. The paper discusses MediaPipe poses on HAR using a Raspberry Pi. The author collects their dataset and reports an accuracy of 96.8% on this dataset.

B. POSE ESTIMATION

The method proposed in [32] for human pose recognition based on Deep Neural Network (DNN) was executed by training 4000 training and 1000 testing images using a DNN based regression model that performed better results on pose estimation as compared to generic CNNs, which were initially designed for classification. The authors in [33] discuss how machines can understand and identify humans and their interactions. Tu et al. [34] present a novel multi-person 3D pose estimation approach. Puwein et al. [35] proposed a method to jointly perform camera pose estimation and human pose analysis from a video recorded by a set of cameras separated by a comprehensive baseline. Mobini et al. [36] discuss the Kinect's skeleton tracking accuracy for upper body rehabilitation applications. Kim et al. [37] present a 3D human pose estimation system from monocular images and root joints and adding joint angle ranges for pose balancing; they developed a complete body humanoid robot model by using the 2D skeleton poses calculated by the ready-made deep learning method, MediaPipe Pose, as the input and fitting through re-projecting the 3D humanoid robot model to the 2D model at the joint angle level using the fast optimization method. The authors in [38] propose a novel approach to estimating human posture by recovering the 2 Dimensional (2D) position of each joint from several photos taken simultaneously at different angles. Kanazawa et al. [39] describe a way to fit a human mesh to an 'in the wild image' of a person. Cao et al. [5] discuss the usage of part affinity fields for multi-person pose estimation. Sun et al. [7] propose a method to preserve the high-resolution representation while estimating pose and compare it to existing methods. It is concluded that the HRNet, a variation of the ResNet, performs better than their previously proposed resnets [40].

C. EXERCISE EVALUATION AND ANALYSIS

Teikari and Pietrusz [41] talk about the state of precision strength, tools, and feedback system in their survey. In [42], the authors propose an exercise evaluation system using BlazePose as a pose estimation algorithm. Taware et al. [43] discuss the working and integration of Artificial Intelligence (AI) in workout assistants and fitness guides, which uses MediaPipe pose estimation to keep track of users' body postures while doing exercises to avoid any injuries. Madanayake et al. [44] discuss a fitness mate and its design process by implementing a system to enable users to perform exercises and avoid injury even when unsupervised. Jain et al. [45] introduce pose trainer, an AI trainer that employs the BlazePose tool from MediaPipe's Pose estimation module to detect a user's posture and then assess the pose of an activity to offer helpful feedback. Saraee et al. in [46], introduce a system called PostureCheck, which assesses the posture of a person exercising using a Microsoft Kinect camera and Bayesian estimation. A plan was presented [47] to implement a CNN model that was trained on Common Object in Content (COCO) for human pose estimation to monitor user workouts. Park in [48] design a mobile personal workout assistant using a deep neural network utilizing about 20000 data points from a squat workout data set. Varghese et al. [49] propose a real-time fitness activity recognition and correction solution using deep neural networks and compare LSTM-RNNs, ConvLSTMs, and Generative Adversarial Networks (GAN)s to classify exercises.

While authors in [51] worked with MediaPipe and You Only Look Once (YOLO)s pose estimation algorithms to propose functions to transform the physical body into Artificial Reality (AR) and Virtual Reality (VR) worlds. Kwon and Kim [52] introduce a system that can correct the posture of a subject in real time using OpenCV and MediaPipe. A 3D pose estimation tool that provides visual feedback to users learning how to perform exercises was designed by [53]. Agarwal et al. [54] study yoga applications that use AI to

motivate their customers with personalized experiences and positive feedback such as voice guidance and reminders. Chaudhari et al. [55] describe a method for correcting a practitioner's posture while practising Yoga Asana that is based on Machine Learning (ML) approaches. Chaudhari et al. [55] use a computer web camera along with a MediaPipe to extract joint angles from 2D to 3D converted data and compare it to joint angles extracted from an expert yoga position. Wang et al. [56] propose a portable quantifiable deadlift evaluation system, which uses Body Sensor Network (BSN) with inertial and surface Electromyography (sEMG) information to extract the human motion information and segment the deadlift into certain phases to realize the detailed analysis.

Oh and Kim [57] use two devices, a Kinect and a Wii Balance board, to capture data and classify squats as correct and incorrect using a SVM and a naive Bayes classifier. Virtucio and Naval [58] developed a support vector machine model that can classify squat postures based on the coordinates of body landmark data extracted through MediaPipe pose. The classifier obtained an accuracy score of 92.92%. Authors in [59] use a mono camera along with MediaPipe to extract 3D data from 2D video data and classify squats using Double Exponential Smoothing (DES) while authors in [60] discuss how to apply pattern recognition and ML techniques to identify whole-body movement patterns during the performance of deep squats and hurdle steps. Rungsawasdisap et al. [61], [62] describe a method to recognize the squat action and classify it into six different types of squats using HMM and CNN, respectively. They use a perception neuron motion capture suit to assign nodes to the body and recover visual or movement data. Ogata et al. [10] discuss a method in which the pose data is extracted using the method described by [39], and the distances between each node are normalized. The technique uses a 1 Dimensional (1D) CNN to classify the squats into seven types. Authors in [1] describe a system, USquat, that utilizes computer vision and ML for understanding and analyzing squats. Ota et al. [63] address the issue of whether OpenPose-based motion analysis has sufficient reliability and validity. They use a model that analyses the motion of a bilateral squat to verify the same and determine which factors are most crucial when analysing squats and the accuracy of the data the model produces. Zhang et al. [64] discuss a squat motion detection model which is designed by combining the MediaPipe algorithm and the You Only Look Once v5 (YOLOv5) network. The result was that the method could effectively detect deep squatting movements, eliminate false detection rates, and improve the algorithm's robustness in complex environments with an accuracy rate of 96%.

D. ACTION IMITATION

Yu and Zou [65] proposed an imitation system for gait, which can recreate gait for some time from one gait cycle or one step. They also implement gait recognition using Hidden Conditional Random Field (HCRF) with SVM classifier.

Chaudhari et al. [66] explored further attention models, including the Luong style and Bhadnau style Attention models. Luong et al. [67] suggested an attention model for NN-based Translation in which Luong style Attention, or dot product attention, was proposed. Hochreiter and Schmidhuber [68] introduced LSTMs, which is a variation of the RNN and retains context for short-term memory. Cho et al. [69] introduce Gated Recurrent Units (GRU), which is a variant of the LSTM.

AI trainer has been becoming popular in fitness and sports performance evaluation as it offers valuable insights and assistance in assessing squat technique, form, and potential areas of improvement. These days, many tend to use online platforms to perform exercises; however, there is no provision for feedback and performance improvement. AI algorithms can analyze video footage of individuals performing squats and provide real-time feedback on their form. Computer vision techniques can detect key joint angles, body posture, and alignment deviations. This information can help identify improper squatting techniques and suggest corrective actions. Pose estimation algorithms can track the positions of body joints in real-time or from recorded videos. This can provide quantitative data on joint angles, alignment, and movement patterns during squats. This information is used to improve and correct the poses of individuals.

AI models can be trained on a large dataset of correct and incorrect squatting techniques. By analyzing new squat videos, these models can classify and score the quality of each squat based on predefined criteria. This can provide standardized and consistent evaluation, reducing subjectivity. AI can simulate biomechanical models of squatting, considering factors like muscle activation, joint forces, and balance. These variables are analyzed to understand the impact of different techniques on the body, aiding in developing safer and more efficient squatting practices. Wearable sensors or devices can provide real-time squat performance data, including depth, speed, and balance. AI can process this data to offer instant feedback to users, helping them adjust their form and technique on the spot. AI-powered systems can analyze an individual's squat performance over time, identify weaknesses or areas needing improvement, and generate personalized training plans. These plans can help individuals gradually enhance their squat technique and overall strength. AI-trainers can analyze squat performance to detect potential injury risks, such as excessive joint stress or poor alignment. AI trainer virtual coaches can guide users through squatting exercises, providing real-time feedback and instructions. This can be particularly useful for individuals who don't have access to in-person trainers. The proposed research introduces and examines the squat performer's predicted limb lengths to evaluate each person fairly and equitably. Our approach combines the benefits of MediaPipe for pose estimation for squat data collected using the stereo camera, Encoder architecture for squat classification, and curve

fitting for squat correction. The main objective of this work includes

- To propose and implement a deep learning architecture to identify if the individual performing the squat exercises to the best of their ability based on their body proportions.
- The study expands the prior work in the Action Quality Assessment (AQA) area to identify and assess squat movement variations.
- To allow beginners to test their understanding of the subject and to monitor their posture and progress during squat sessions. This research will also aid the community in assisting users in understanding their form and performing the proper squat.
- To propose a cost-effective system that identifies squats and corrects incorrect squats by displaying the correct form to the person using it.

The structure of the paper is organised as follows. SectionII-A presents the proposed methodology describing the dataset collection, stereo vision, pose estimation, classification and correction method with a detailed explanation of each sub-blocks involved, followed by result analysis and discussion in sectionIII. Finally, the conclusion is presented in the section IV.

II. PROPOSED METHODOLOGY

A. PROBLEM DEFINITION

Beginners in the gym tend to make mistakes in exercises that may lead to lifelong injuries. This is true for more experienced athletes as well. One slight imprecision in an essential activity, such as the squat, can cause the whole form of the person to vary. The squat is one of the most nuanced exercises, and multiple myths about it cause beginners to shy away. Hence, they make common mistakes such as not allowing their knees to cross their toes, not squatting deeper than parallel, etc. This model aims to aid beginners in correcting their form and squat, using advanced tools such as deep learning and computer vision, to the best of their ability.

Fitness trainers or coaches have become a common practice. However, it is costly to employ a coach and thus only sometimes feasible for a beginner in the gym. Using pre-existing software and equipment, this model attempts to alleviate the problem by suggesting a method to help beginners get introduced to fitness and working out by assisting them to practice better and avoid injuries while performing exercises by focusing on the squat. Implementing this work may also assist sports scientists and physiotherapists analyse, diagnose and treat people more efficiently. The contribution of the work is as follows: (i) The video dataset comprising 1332 is collected on seven types of squats using more than 50 volunteers (ii) The stereo camera setting has been done to collect the data that is fed to MediaPipe to estimate the pose (iii) The classifier is trained to classify the type of squat. (iv) The wrongly performed squat is identified, and a corrected version is shown to the user (v) A comparative study has been performed to compare the proposed method with the existing state-of-the-art methods.

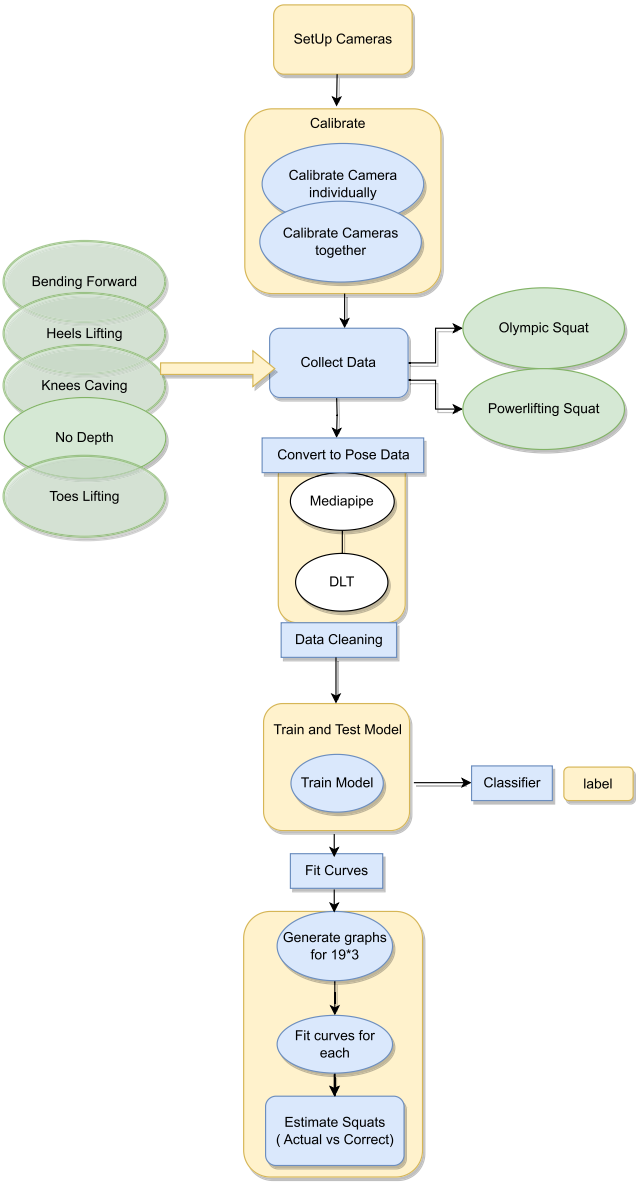


FIGURE 1. Process flow of the proposed approach for squat detection and correction.

The flowchart of the proposed approach for squat analysis is depicted in Fig. 1. After calibrating both cameras, the stereo camera setup collects the data. The dataset is then labelled to train the deep learning model. Initially, the pose is estimated by feeding the data through MediaPipe, followed by data cleaning. Data is split into train and test data to train a custom auto-encoder model. The model is then used to classify the squat type. The corrected version of the squat is shown to the user along with the wrong form that helps to correct the squat. This system continuously improves the performance of squats. Each sub block of Fig. 1 is explained in the following sections.

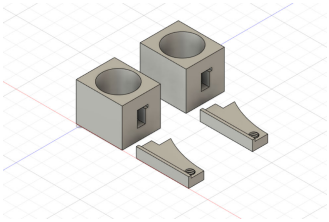


FIGURE 2. 3D model for camera holder.

(a) Intel depth dense camera 0

(b) Camera holder with Intel depth sense cameras 0 and 1

FIGURE 3. Proposed experimental setup to collect the dataset.

(a) Camera 1 (b) Camera Holder Top view (c) Camera setup side view

FIGURE 4. Proposed experimental setup.

(a) Stereo Calibration Frame for Camera 0 (b) Stereo Calibration Frame for Camera 1 (c) Mono Calibration Frame for Camera 0

FIGURE 5. Images used for calibration process.

B. STEREO VISION MODULE

Stereo vision is extracting 3D information from digital images taken by two cameras separated horizontally to obtain two different views of the same scene. Stereo vision is employed in this project to get more accurate pose information. A pair of Intel Real Sense D435 depth cameras are used (however, one can use any RGB camera for this purpose). The color images obtained from both cameras are used as input to determine the pose of a person [70], [71].

An experimental setup was made to collect the dataset for the proposed work. The mounts for the cameras were 3D printed along with the mounting screws, as shown in Fig. 2. The experimental setup is depicted in Fig. 3-4. The cameras are mounted on a set of Poly Vinyl Chloride (PVC) pipes held 50× 60 cm apart. This is to keep the angle subtended at the intersection of the focal points between 30 and 60 degrees. The mount places the camera 40 cm off the base of the setup, effectively placing the camera 1.2 m above the ground. This allows a full view of the person performing the squat, including the person's feet and head.

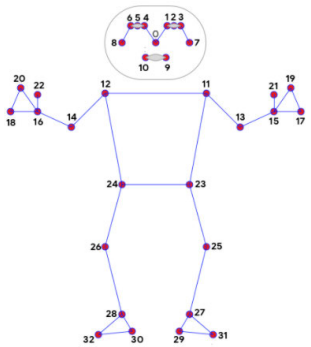
1) CALIBRATION

Calibration is used for configuring an instrument to provide a result for a sample within an acceptable range. Calibration of the cameras is done to obtain the intrinsic and extrinsic parameters so that a proper estimate of the actual limb lengths may be obtained while estimating pose. A code derived from python-stereo-camera-calibrate [72] is used to calibrate the cameras. A chessboard pattern printed on paper is used to calibrate the cameras. While calibrating, it is necessary to obtain Root Mean Square Error (RMSE) of less than or equal to 0.3 for individual cameras and less than or equal to 0.5 for both cameras. Sample images are shown in Fig. 5

2) STEREO POSE ESTIMATION

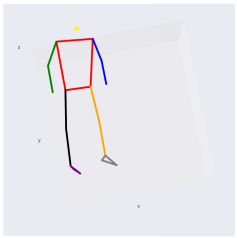
Both cameras record videos of the same scene simultaneously. The dataset is collected and converted to pose data using a code derived from bodypose3d [73]. MediaPipe is employed to estimate the pose due to the following reasons: (i) It is more or less robust to occlusions, including self-occlusion; (ii) It can estimate the pose of a single person in a frame; (iii) It is computationally inexpensive and has a low runtime; (iv) It is readily available and is an off-the-shelf algorithm, used in research regularly.

Pose data is then extracted from both videos and Direct Linear Transform (DLT) is used to transform 2D pose data into 3D pose data. DLT is a powerful technique in machine vision that enables 3D reconstruction and camera calibration. A 2-camera setup, DLT, can improve accuracy and obtain depth perception. DLT works by establishing correspondences between 3D points in the real world and



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

(a) MediaPipe's estimated key points



(b) Skeleton from bodypose3d

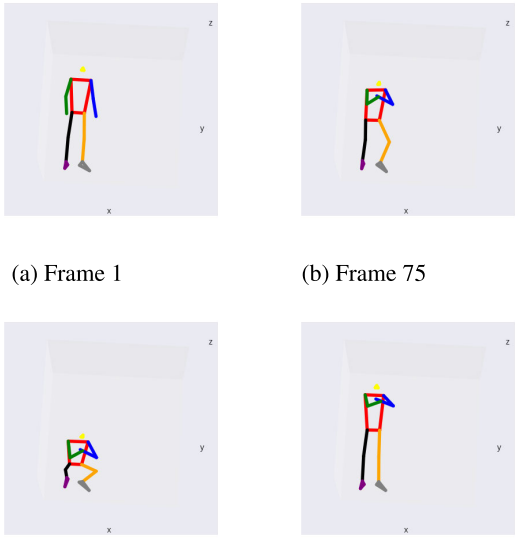
FIGURE 6. MediaPipe's application to extract the skeleton of the body.

their corresponding 2D projections in the image plane. The DLT method is based on Singular Value Decomposition (SVD)). With a two-camera setup, two cameras capture the same scene from different viewpoints, providing additional information for depth estimation. Since the DLT method provides us with a system of two equations for one camera view, a pair of cameras (or more) are required to estimate the world coordinates.

Initially, MediaPipe extracts landmarks to convert videos into pose data. Further, the camera parameters and DLT estimate the world coordinates. The pose data is proportional to the person in the video compared to MediaPipe's normalized data. MediaPipe extracts thirty-three key points, of which 19 are extracted from the video data using this algorithm. These key points can be shown in Fig. 6. Fig 7 depicts sample frames of the skeleton and key points obtained after passing through the GLT algorithm.

C. DATASET COLLECTION

More than 50 individuals volunteered to assist in data collection. All of them were asked to sign a form stating that they consented to the data collection. The videos are recorded in different environment settings, indoor and outdoor scenes. A wide range of lighting conditions were selected under various clothing options. Each data span around 2 to 5 seconds. The length of the video varies based on the individual, and the distribution of frame length over the dataset is depicted in Fig. 9. This resulted in a dataset comprising over 1332 pieces of video data spanning seven different classes. The distribution of the dataset over these



(a) Frame 1 (b) Frame 75

(c) Frame 150 (d) Frame 275

FIGURE 7. Sample frames from the pose estimation video.

TABLE 1. The details of squat dataset.

Dataset	
Class	Number of Datum
Bending forward	163
Heels Lifting	177
Knees Caving	152
No Depth	182
Toes Lifting	154
Olympic Squat	264
Powerlifting Squat	240
Total	**1332**

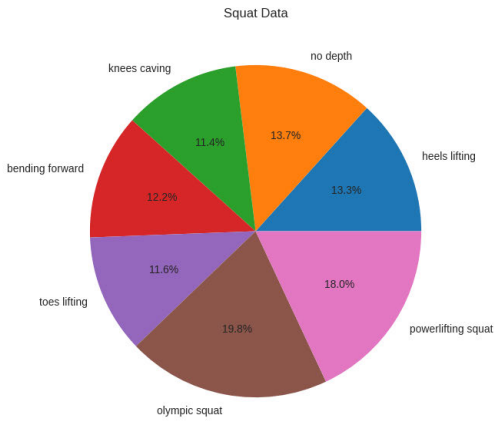


FIGURE 8. Pie chart of the data collected.

classes is presented in the Pie chart, Fig.8. The classes and the corresponding number of data points are shown in Table 1. This approach did not employ data augmentation, as it may change the exercise form. Thus, vertical augmentation makes sense in this method. Each volunteer is asked to perform good and bad squats of different types.

The length of the data varied between 30 frames and 291 frames. Hence, it was optimal to pad the data to

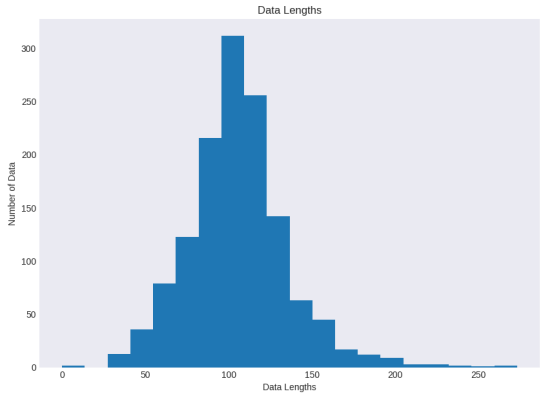


FIGURE 9. Histogram of the data lengths.

300 frames. The padding value selected was '-1'. The variation in the data length is shown in Fig. 9. The data is split in the 80:20 ratio, where 80% of the dataset is used for training the model, and 20% is used for validation. We have used 1332 videos of which 1065 are used to train the models, and the rest 267 are employed to validate the data. We have tested the algorithm on real-time videos and provided the corrected form in case the user fails to perform a correct squat. The evaluations are performed using the accuracy and loss curves for the datasets. The datasets are manually labelled as good and bad squat by an expert before training the model, which is used to measure the model performance.

D. PROPOSED CLASSIFIER FOR SQUAT CLASSIFICATION

To classify squats, the proposed custom autoencoder model was generated and depicted in the block diagram Fig 10.

1) CLASSIFIER MODEL

A stacked Bi-GRU model, with an attention layer, was chosen for classification. GRUs are used since they can provide context to time series data. Using a bidirectional network allows context to flow in both directions, forward and backward.

Below is a brief explanation of the GRU layer.

$$h_t = GRU_{enc}(x_t, h_{t-1}) \tag{1}$$
$$s_t = GRU_{dec}(y_t, s_{t-1}) \tag{2}$$

In Equation 1 and Equation 2, $\boldsymbol{h_t}$ refers to the encoder hidden state at time-step $\boldsymbol{t}$, with input token embedding $\boldsymbol{x_t}$ and $\boldsymbol{s_t}$ refers to the decoder hidden state at time-step $\boldsymbol{t}$, with input token embedding $\boldsymbol{y_t}$.

An attention mechanism is used in the model. It allows the model to focus on specific input parts by assigning weights to different positions. Attention mimics the human behaviour of focusing on essential things while ignoring the less relevant ones. Attention has been used with Recurrent Neural Networks and their variants with great success. In this model, Luong style attention or Dot Product Attention is used.

TABLE 2. Parameters of the proposed model.

Model	
Parameter	Value
Optimizer	'adam'
Learning Rate	0.001
Epochs	1000
Batch Size	128
Callbacks	'model_checkpoint'
Metrics	'categorical_accuracy'
Loss	'categorical_crossentropy'

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{3}$$
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{4}$$
$$e_{ij} = a(s_{i-1}, h_j) \tag{5}$$

In Equation 3, Equation 4 and Equation 5, α_{ij} refers to the attention weight with i^{th} decoder step and j^{th} encoder step, resulting in context vector $\mathbf{c_i}$.

If $\boldsymbol{h_t}$ and $\boldsymbol{s_t}$ have the same number of dimensions, then:

$$e_{ij} = s_{i-1}^T h_j \tag{6}$$

Otherwise:

$$e_{ij} = s_{i-1}^T W h_j \tag{7}$$

Finally, output $\boldsymbol{o_i}$ is produced by:

$$s_t = tanh(W[s_{t-1}; y_t; c_t]) \tag{8}$$
$$o_t = softmax(Vs_t) \tag{9}$$

The model details are given in Table 2.

The model is trained for 1000 epochs, with a learning rate 0.001. The model also has dropout layers that prevent overfitting. The best-trained model is selected, which has a validation accuracy of 94%.

2) ESTIMATION OF GOOD SQUATS

The proposed method classifies the given squat as good or bad. A set of landmarks is extracted through MediaPipe. A curve is fitted to each coordinate of each key point on the body for a good squat video.

$$y = k_0 + k_1 x + k_2 x^2 + k_3 x^3 + \ldots.. + k_{n-1} x^{n-1} + k_n x^n \tag{10}$$

In Equation 10, k_0 refers to the y intercept of the curve. After observing the curves of multiple coordinates of multiple squats, it was concluded that the shape of the curve remains the same. At the same time, the y intercept varies, and the curve is stretched over a larger timeframe.

Hence, a curve was fitted to one good squat, then shifted up or down about the y-axis using the y-intercept. The curve is also stretched or squeezed over a timeframe using time

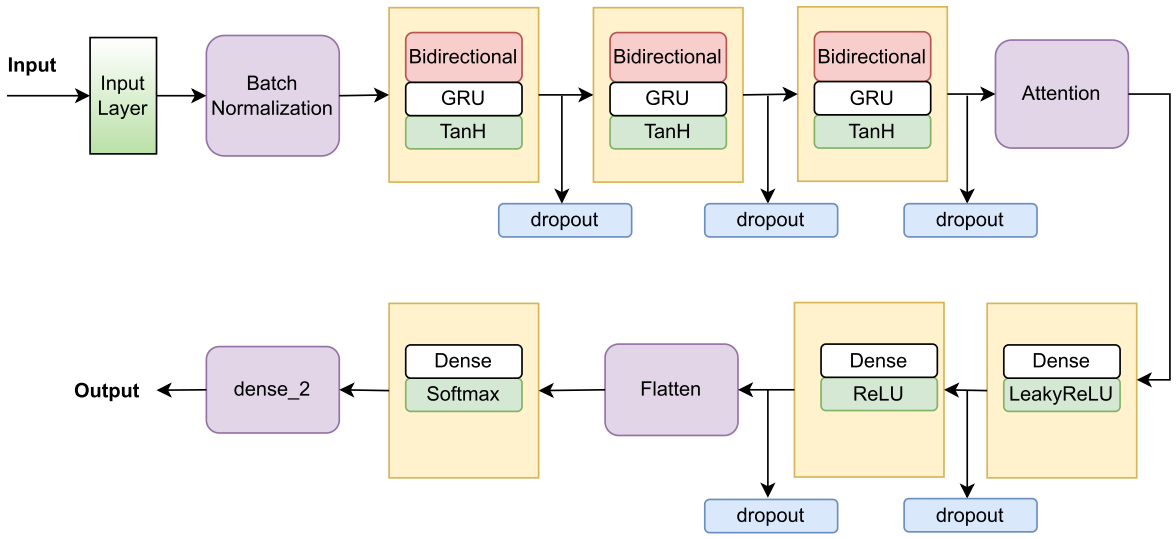


FIGURE 10. The proposed architecture of Bi-GRU with attention for squat classification.

TABLE 3. Qualitative analysis of existing models for squat classification.

Comparative Study			
Name	Algorithm	Accuracy	Issues
AI Fitness Trainer using MediaPipe [2]	Hard-coded	Poor	Does not have multiple classes, does not take into consideration nuances
squatevaluation [4]	3D CNN	Poor	Has only two classifications, tends to classify bad squats as good squats, computationally expensive for larger sets of data
SquatClassificationAndCounting [6]	3D CNN	Better	Has multiple classes, but they are based only on depth, computationally expensive for pose estimation
Standard-Squat-Posture-Classifier [8]	SVM	Better	Real-time classifier gives multiple classifications for a squat.
IVU [9]	LSTM	Better	Would not accept video length that is not equal to stride, hence gave multiple classifications for a single video, no option to pad data to length
squatDepth [3]	Hard-coded	Poor	Does not have multiple classes, does not take into consideration nuances

scaling. This is done to extract the estimated good squat. This preserves the limb lengths of the person doing the squat while evaluating the squat.

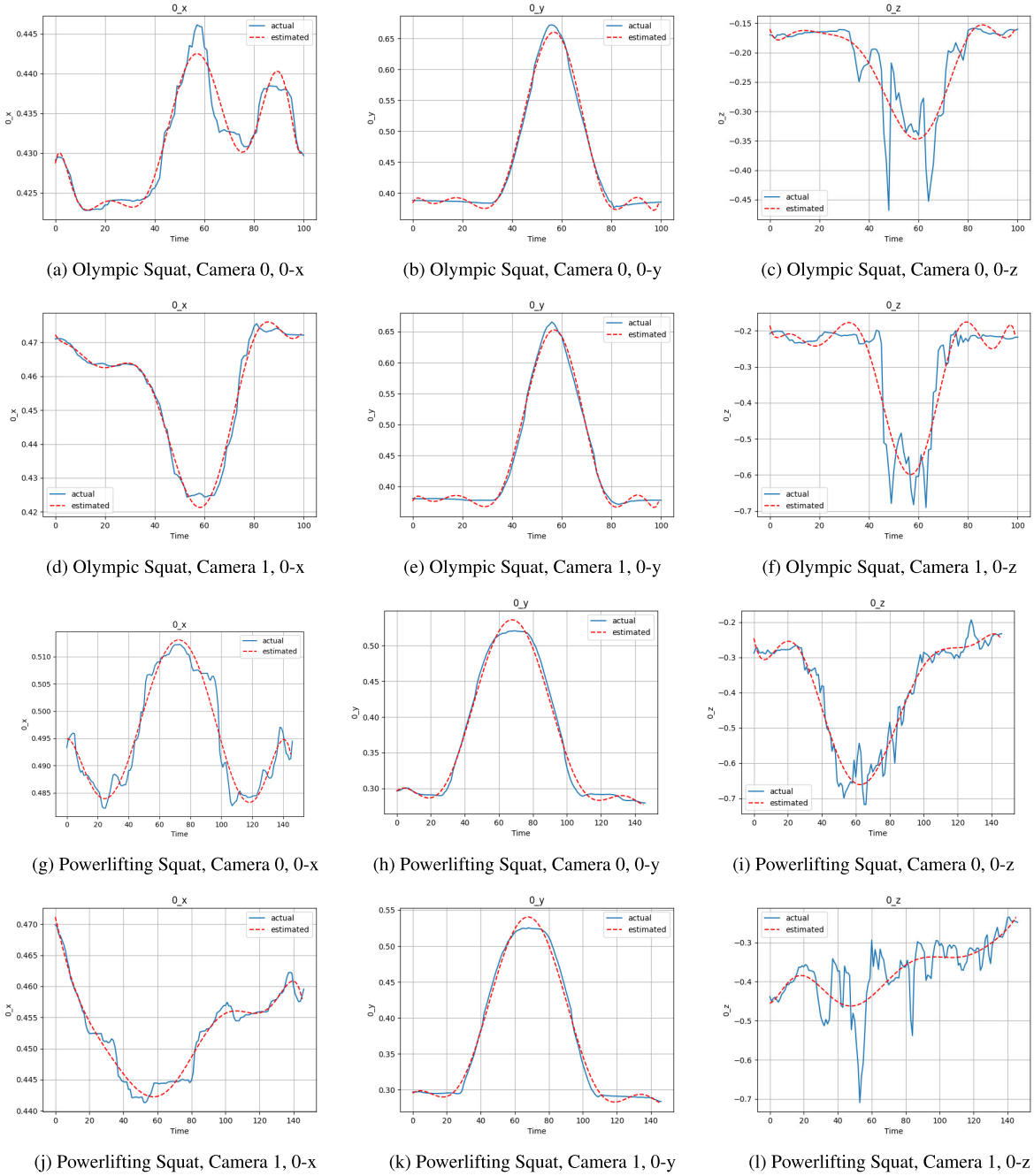
- Fig. 11a - 11c show the curves fit for the x, y and z co-ordinates of the 0th key-point of camera 0 for an Olympic squat.
- Fig. 11d - 11f show the curves fit for the x, y and z co-ordinates of the 0th key-point of camera 1 for an Olympic squat.
- Fig. 11g - 11i show the curves fit for the x, y and z co-ordinates of the 0th key-point of camera 0 for a powerlifting squat.
- Fig. 11j - 11l show the curves fit for the x, y and z co-ordinates of the 0th key-point of camera 1 for a powerlifting squat.

Fig. 12 shows the 3d curves representing the squat over video frames for the Olympic and powerlifting squat. This explains how a person performs the squat in a given time frame and each camera.

III. RESULTS AND DISCUSSION

Due to changes in the algorithm and dataset, we found the difficulty to compare the proposed method with existing squat analysis approaches. Hence, in this section, we compare the existing approaches on the particular dataset mentioned by the authors in respective papers. We collected the dataset required for squat classification from various sources. It was concluded that existing codes should be evaluated on a dataset prescribed to that particular approach. The dataset is then labelled and split into videos of single squats rather than sets of squats. This helped to evaluate the existing models better. The set of datasets is as follows: (i) Countix Dataset (ii) Kinetics 700 Dataset (iii) MultiModal - Fit (MM - Fit) Dataset (iv) UCF-101 Dataset (v) Singe Individual Dataset from Temporal Distance Matrices for Squat Classification (vi) FitnessAQA Dataset

We conducted a comparative study on the available dataset using state-of-the-art methods. We analyse each process with its advantages and disadvantages. We considered six



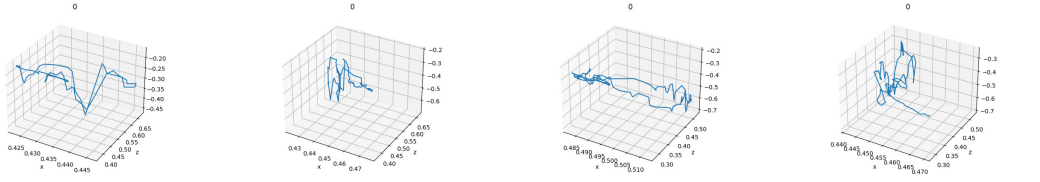
(a) Olympic Squat, Camera 0, 0-x (b) Olympic Squat, Camera 0, 0-y (c) Olympic Squat, Camera 0, 0-z

(d) Olympic Squat, Camera 1, 0-x (e) Olympic Squat, Camera 1, 0-y (f) Olympic Squat, Camera 1, 0-z

(g) Powerlifting Squat, Camera 0, 0-x (h) Powerlifting Squat, Camera 0, 0-y (i) Powerlifting Squat, Camera 0, 0-z

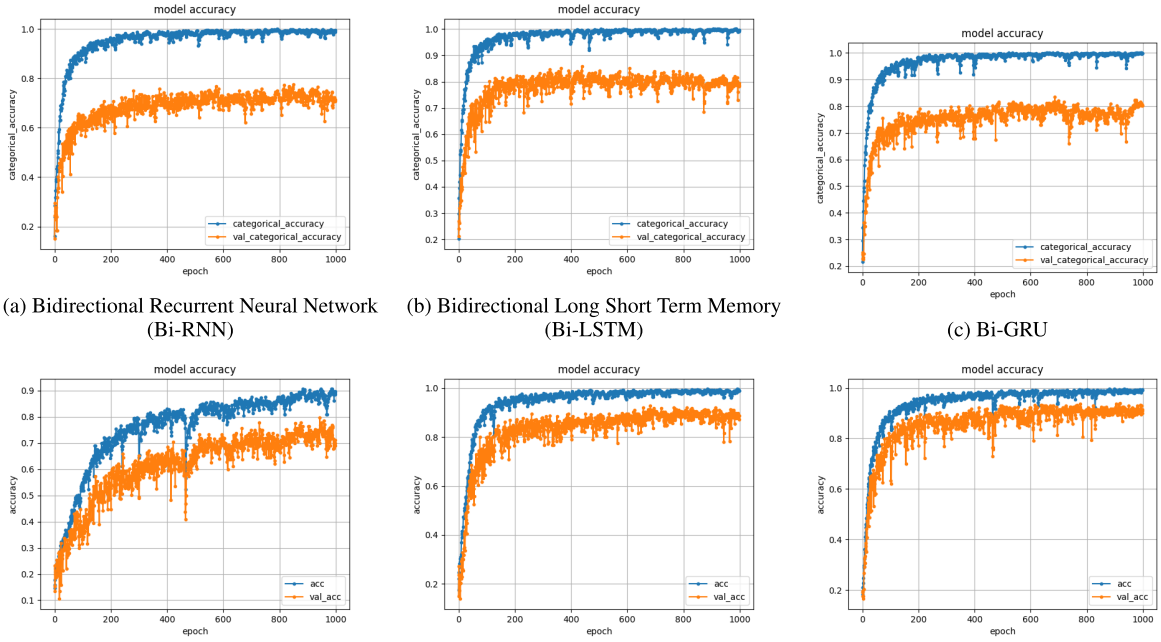
(j) Powerlifting Squat, Camera 1, 0-x (k) Powerlifting Squat, Camera 1, 0-y (l) Powerlifting Squat, Camera 1, 0-z

FIGURE 11. **Sample images of the curve fit to estimate the correct squat.**



(a) Olympic Squat, Camera 0, 0 (b) Olympic Squat, Camera 1, 0 (c) Powerlifting Squat, Camera 0, 0 (d) Powerlifting Squat, Camera 1, 0

FIGURE 12. **Tracking of a body position over time.**



(a) Bidirectional Recurrent Neural Network (Bi-RNN)

(b) Bidirectional Long Short Term Memory (Bi-LSTM)

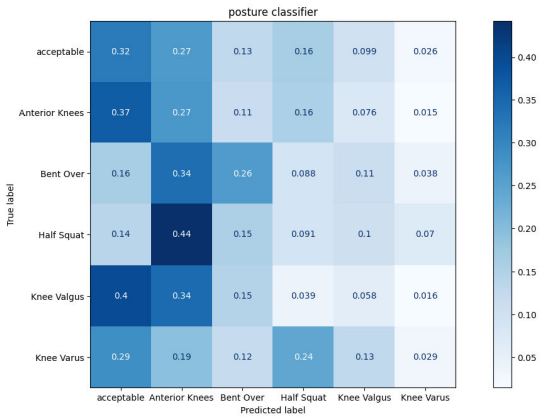
(c) Bi-GRU

(d) Bi-RNN with Attention

(e) Bi-LSTM with Attention

(f) Bi-GRU with Attention

FIGURE 13. **Accuracy versus Epoch of the models.**



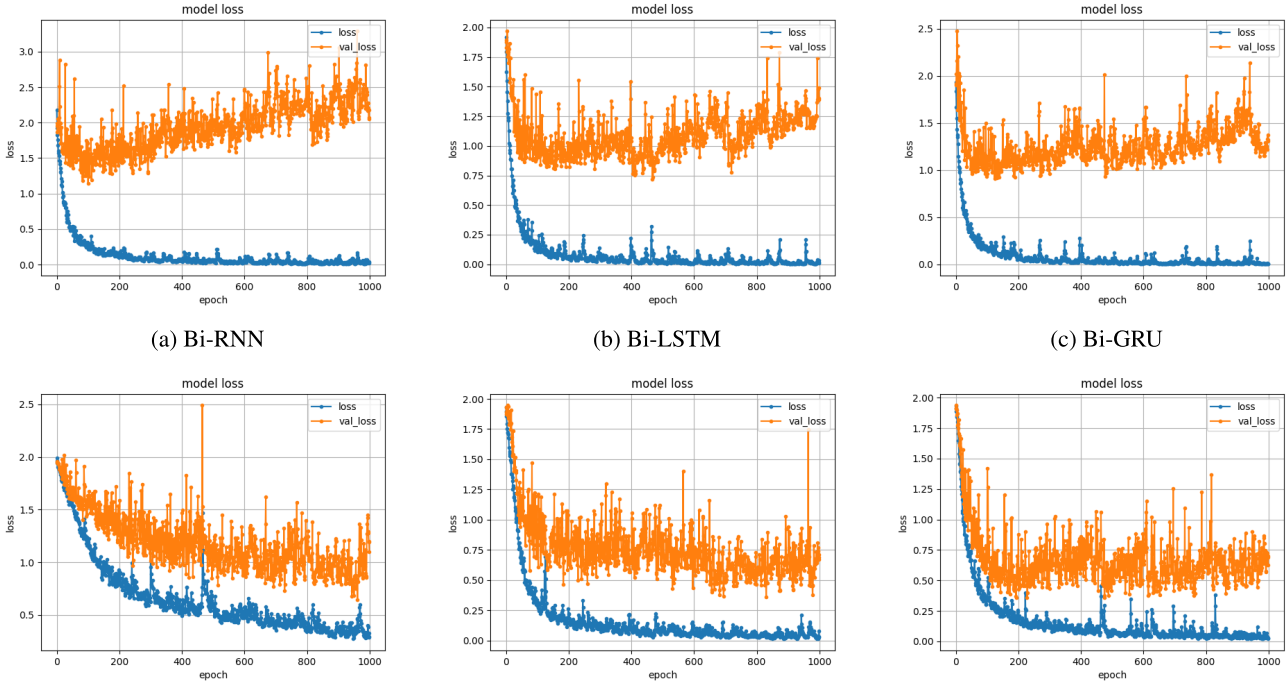
(a) Standard-Squat-Posture-Classifier [8] Confusion Matrix

FIGURE 14. **Confusion matrices from the comparative study.**

squat analysis algorithms to examine both quantitatively and qualitatively. A qualitative analysis is provided in Table 3, and confusion matrices are depicted in Fig. 14. It has been observed that the existing solutions use different datasets and different approaches. Thus, it is unfair to compare these algorithms. Hence, we propose a novel approach to make it more general by collecting our dataset using a stereo camera and deep learning techniques to analyse the squat. We will make our dataset and algorithm publicly available for researchers in the future.

A comparison was made between different models, and it was concluded that the Bi-GRU model with an attention layer performed best. Table 4 shows the overall accuracies of the models. The Fig 13 explains the model performance for Bi-RNN, Bi-LSTM, and Bi-GRU with and without attention layer, respectively. The Bi-GRU with attention layers performs better for the current dataset, proving high accuracy for training and validation datasets. Similarly, the loss graphs of each model are presented in Fig 15. The confusion matrix provides the details of true positive, false positive, true negative, and false negative for all selected classes. Fig. 16 depicts the confusion matrix for tested models. Table 5 shows the class-specific accuracies for the models. The Labels are written in short in the table; the expanded forms are given below:

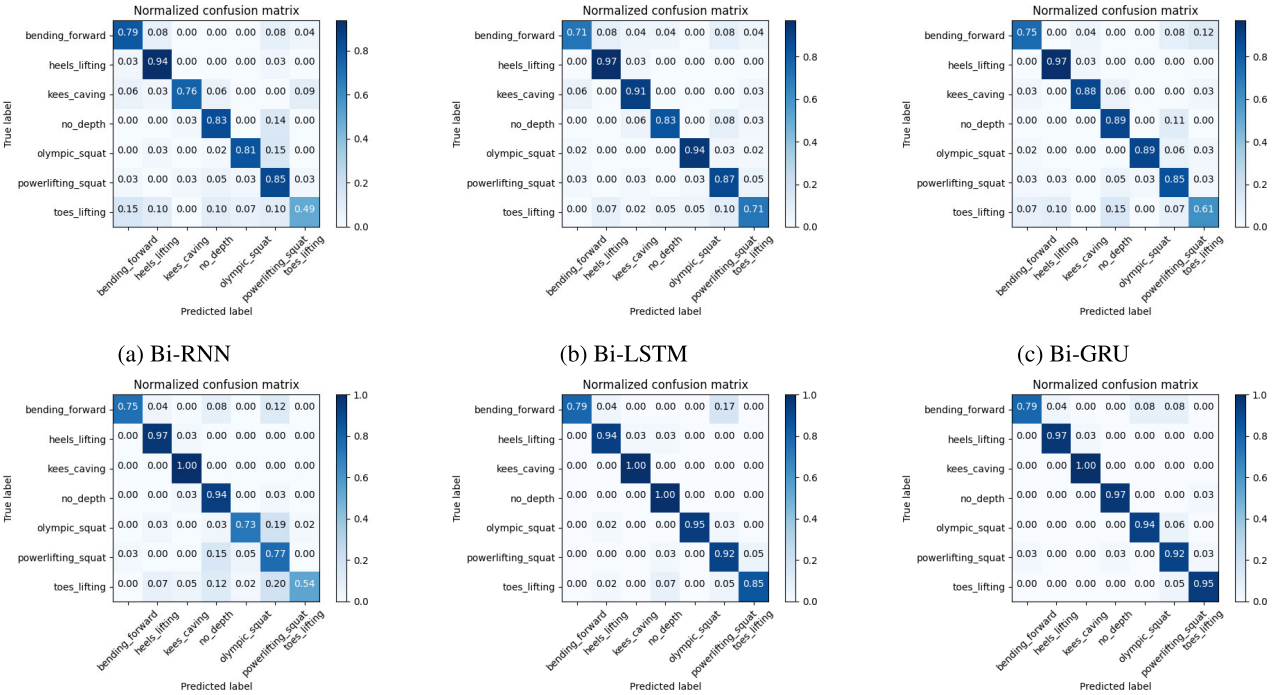
From Tables 4 and 5, we can draw the following inferences: adding the attention layer improves the overall accuracy of the models. Providing context for short-term memory via long-term memory enhances the performance of models. The LSTM model outperforms the GRU model without an attention layer, but the GRU model outperforms the LSTM model with an attention layer. The RNN model does not gain much performance even after adding an attention layer. The accuracy is low in both groups. All models underperform on Bending forward (BF). This may be because of under-representation in the dataset, due to the predisposition of the model to have this fault, or due to higher portions of unclean data being present in this class. The models without the attention layer underperform on Toes Lifting (TL), but those with attention perform well on TL. Bi-GRU with an attention layer provides highest accuracy of 94% among the state-of-the-art classification models and is highlighted in bold. This is followed by Bidirectional Long Short Term Memory (Bi-LSTM) model



(a) Bi-RNN (b) Bi-LSTM (c) Bi-GRU

(d) Bi-RNN with Attention (e) Bi-LSTM with Attention (f) Bi-GRU with Attention

FIGURE 15. **Loss versus epoch of the models.**



(a) Bi-RNN (b) Bi-LSTM (c) Bi-GRU

(d) Bi-RNN with Attention (e) Bi-LSTM with Attention (f) Bi-GRU with Attention

FIGURE 16. **Normalized confusion matrices of the models.**

with attention layer. However, the Bidirectional Recurrent Neural Network (Bi-RNN) model provides least accuracy of 77.5% for our dataset. The normalized confusion matrices reflect the accuracy of the proposed architecture.

The current research could be further extended to exercise analysis using ML algorithms and imitating human movement using the same. In addition, it can be extended to yoga and other type of exercises as well. In addition, several

TABLE 4. Comparison of different models.

Model Comparison	
Model Name	Accuracy
Bi-RNN	77.52%
Bi-LSTM	85.76%
Bi-GRU	83.52%
Bi-RNN with Attention	79.77%
Bi-LSTM with Attention	92.88%
Bi-GRU with Attention	**94.00%**

TABLE 5. Comparison of accuracies of different models for different squat types.

Accuracy Comparison							
Model Name	BF	HL	KC	ND	OS	PS	TL
Bi-RNN	**0.79**	0.94	0.76	0.83	0.81	0.85	0.49
Bi-LSTM	0.71	**0.97**	0.91	0.83	0.94	0.87	0.71
Bi-GRU	0.75	**0.97**	0.88	0.89	0.89	0.85	0.61
Bi-RNN-Attention	0.75	**0.97**	**1.00**	0.94	0.73	0.77	0.54
Bi-LSTM-Attention	**0.79**	0.94	**1.00**	**1.00**	**0.95**	**0.92**	0.85
Bi-GRU-Attention	**0.79**	**0.97**	**1.00**	0.97	0.94	**0.92**	**0.95**

Bending forward (BF). Heels Lifting (HL), Knees Caving (KC), No Depth (ND), Olympic Squat (OS), Powerlifting Squat (PS), Toes Lifting (TL)

classification algorithms could be tested. A few such approaches would be: using different types of LSTM networks such as GCA-LSTM, ST-LSTM, FC-LSTM, ConvLSTM, etc. for classification purposes. It is helpful to build Attention models customized to the task and implemented for the classification and estimation tasks. The proposed method used the curve fitting method to generate an excellent squat. However, using NN based regressor on developing the excellent squat could be explored. Implementing different pose estimation algorithms such as OpenPose, VoxelPose, YOLOv7 Pose, etc., can open new room for improvement for the task of exercise analysis. Using multi-view systems, with views that can see the subject from all sides for HAR having more classes in the data, such as buttwink, knee varus, lack of control on the eccentric, etc., would diversify the squat data.

IV. CONCLUSION

The paper proposes a novel method to classify various squat types and recommends the right squat version. A dataset comprising 1332 individual records spanning seven different classes was collected as part of the study. We employed a custom Bi-RNN, Bi-LSTM, Bi-GRU architecture with and without attention layer for squat classification. A stacked bidirectional GRU classifier with an attention layer outperformed among selected techniques to classify the squat types. The Bi-GRU classifier reported an accuracy of 94%, which showed the best among existing research in this area. A comparison has been conducted with state-of-the-art models, which proved that the stacked Bi-GRU with attention performed consistently best. Finally, an estimator was created to give feedback on the inputted squat.

AI systems can track an individual's squat performance over time, highlighting improvements and areas that still need work. This can help users stay motivated and focused on their fitness goals. As with any application of AI, it's important to ensure that the data used for training and evaluation is diverse, representative, and of high quality. Balancing technological assistance and human coaching is crucial for effective and safe squat evaluation.

ACKNOWLEDGMENT

The authors thank the volunteers who participated in the squat data collection process.

REFERENCES

[1] D. N. T. Pham, "USquat: A detective and corrective exercise assistant using computer vision and machine learning," M.S. thesis, Dept. Comput. Sci., California State Univ., East Bay, Hayward, CA, USA, 2021.

[2] LearnOpenCV. *AI Fitness Trainer Using MediaPipe*. Accessed: Mar. 28, 2023. [Online]. Available: https://learnopencv.com/ai-fitness-trainer-using-MediaPipe/

[3] Imarsh64. *Squat Depth*. Accessed: Mar. 28, 2023. [Online]. Available: https://github.com/imarsh64/squatDepth

[4] DNLY09. *Squat Evaluation*. Accessed: Mar. 28, 2023. [Online]. Available: https://github.com/dnly09/squatevaluation

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[6] IKYHCS. *Squat Classification and Counting*. Accessed: Mar. 28, 2023. [Online]. Available: https://github.com/ikyhcs/SquatClassification AndCounting

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. CVPR*, Jun. 2019, pp. 5686–5696.

[8] Renzo Virtucio. *Standard-Squat-Posture-Classifier*. [Online]. Available: https://github.com/renzovirtucio/Standard-Squat-Posture-Classifier

[9] Cypherics. *IVU*. Accessed: Mar. 28, 2023. [Online]. Available: https://github.com/cypherics/IVU

[10] R. Ogata, E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Temporal distance matrices for squat classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2533–2542.

[11] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," in *Proc. VSSN*, Oct. 2006, pp. 171–178.

[12] A. Dubois and F. Charpillet, "Human activities recognition with RGB-depth camera using HMM," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 4666–4669.

[13] P. T. Hai and H. H. Kha, "An efficient star skeleton extraction for human action recognition using hidden Markov models," in *Proc. IEEE 6th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2016, pp. 351–356.

[14] D. Singh, A. K. Yadav, and V. Kumar, "Human activity tracking using star skeleton and activity recognition using hmms and neural network," *Int. J. Sci. Res. Publications*, vol. 4, no. 5, May 2014.

[15] T. Hachaj and M. R. Ogiela, "Human actions recognition on multimedia hardware using angle-based and coordinate-based features and multivariate continuous hidden Markov model classifier," *Multimedia Tools Appl.*, vol. 75, no. 23, pp. 16265–16285, Dec. 2016.

[16] T. Callens, T. van der Have, S. V. Rossom, J. De Schutter, and E. Aertbeliën, "A framework for recognition and prediction of human motions in human–robot collaboration using probabilistic motion models," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5151–5158, Oct. 2020.

[17] A. Madabhushi and J. K. Aggarwal, "A Bayesian approach to human activity recognition," in *Proc. 2nd IEEE Workshop Vis. Surveill.*, Jun. 1999, pp. 25–32.

[18] K.-T. Song and W.-J. Chen, "Human activity recognition using a mobile camera," in *Proc. 8th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, Nov. 2011, pp. 3–8.

[19] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[21] N. Paulose, M. Muthukumar, S. Swathi, and M. Vignesh, "Recurrent neural network for human action recognition using star skeletonization," *Int. Res. J. Eng. Technol.*, vol. 6, no. 3, pp. 123–130, 2019.

[22] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE Access*, vol. 6, pp. 17913–17922, 2018.

[23] I. A. Putra, O. D. Nurhayati, and D. Eridani, "Human action recognition (HAR) classification using mediapipe and long short-term memory (LSTM)," *TEKNIK*, vol. 43, no. 2, pp. 151–162, 2022.

[24] K. M. Adarsh, U. B. Nagesh, A. V. Doddagoudra, and M. K. Bhat, "Human action recognition using deep learning technique."

[25] N. S. Yadav, S. Ramasubbareddy, and M. Ravikanth, "Neural network-based activity recognition system," in *Innovations in Computer Science and Engineering*. Springer, 2022, pp. 511–518.

[26] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, Jul. 2020.

[27] Z. Zhang, Z. Lv, C. Gan, and Q. Zhu, "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions," *Neurocomputing*, vol. 410, pp. 304–316, Oct. 2020.

[28] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Computer Vision—ECCV 2016*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 816–833.

[29] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[30] H. Jain and G. Harit, "An unsupervised sequence-to-sequence autoencoder based human action scoring model," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.

[31] T.-N. Doan, "An efficient patient activity recognition using LSTM network and high-fidelity body pose tracking," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 8, pp. 226–233, 2022, doi: 10.14569/IJACSA.2022.0130827.

[32] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[33] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[34] H. Tu, C. Wang, and W. Zeng, "VoxelPose: Towards multi-camera 3D human pose estimation in wild environment," in *Computer Vision—ECCV 2020*. Glasgow, U.K.: Springer, Aug. 2020, pp. 197–212.

[35] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys, "Joint camera pose estimation and 3D human pose estimation in a multi-camera setup," in *Computer Vision—ACCV 2014*. Singapore: Springer, Nov. 2015, pp. 473–487.

[36] A. Mobini, S. Behzadipour, and M. Saadat Foumani, "Accuracy of kinect's skeleton tracking for upper body rehabilitation applications," *Disab. Rehabil., Assistive Technol.*, vol. 9, no. 4, pp. 344–352, Jul. 2014.

[37] J.-W. Kim, J.-Y. Choi, E.-J. Ha, and J.-H. Choi, "Human pose estimation using MediaPipe pose and optimization method based on a humanoid model," *Appl. Sci.*, vol. 13, no. 4, p. 2700, Feb. 2023.

[38] S. Sarkar, Y. Jang, and I. Jeong, "Multi-camera-based 3D human pose estimation for close-proximity human-robot collaboration in construction," in *Proc. 9th Int. Conf. Construct. Eng. Project Manag. (ICCEPM)*, Las Vegas, NV, USA, Jun.

[39] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7122–7131.

[40] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 466–481.

[41] P. Teikari and A. Pietrusz, "Precision strength training: Data-driven articial intelligence approach to strength and conditioning," *SportRxiv*, May 2021, doi: 10.31236/osf.io/w734a.

[42] S. Krishnamoorthy, "Identification and analysis of human physical exercise postures using computer vision and deep learning," Ph.D. dissertation, Dept. Ind. Control Eng., Universiti Malaya, Kuala Lumpur, Malaysia, 2021.

[43] G. Taware, R. Kharat, P. Dhende, P. Jondhalekar, and R. Agrawal, "AI-based workout assistant and fitness guide," in *Proc. 6th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA*, Aug. 2022, pp. 1–4.

[44] P. S. Madanayake, W. A. D. K. Wickramasinghe, H. P. Liyanarachchi, H. M. D. M. Herath, A. Karunasena, and T. D. Perera, "Fitness mate: Intelligent workout assistant using motion detection," in *Proc. IEEE Int. Conf. Inf. Autom. Sustainability (ICIAfS)*, Dec. 2016, pp. 1–5.

[45] K. Jain, J. Jadav, M. Yadav, and D. Y. Mane, "AI fitness trainer," *Int. J. Emerg. Technol. Innov. Res.*, vol. 9, pp. 380–385, Apr. 2022. [Online]. Available: https://www.jetir.org/view?paper=JETIR2204658

[46] E. Saraee, S. Singh, A. Joshi, and M. Betke, "PostureCheck: Posture modeling for exercise assessment using the Microsoft kinect," in *Proc. Int. Conf. Virtual Rehabil. (ICVR)*, Jun. 2017, pp. 1–2.

[47] A. Nagarkoti, R. Teotia, A. K. Mahale, and P. K. Das, "Realtime indoor workout analysis using machine learning & computer vision," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 1440–1443.

[48] G. Park, "Design of mobile personal workout assistant using deep learning," Ph.D. dissertation, Singapore Nat. Univ., Singapore, Aug. 2020. Accessed: Mar. 24, 2023.

[49] M. M. Varghese, S. Ramesh, S. Kadham, V. M. Dhruthi, and P. Kanwal, "Real-time fitness activity recognition and correction using deep neural networks," in *Proc. 57th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2023, pp. 1–6.

[50] Rajiv, "Academic project squat classification," Ph.D. dissertation, Univ. Hyderabad, Hyderabad, India, 2022.

[51] Y. Zhang, "Applications of Google MediaPipe pose estimation using a single camera," Dept. Comput. Sci., California State Polytech. Univ., Pomona, CA, USA, 2022.

[52] Y. Kwon and D. Kim, "Real-time workout posture correction using OpenCV and MediaPipe," *J. Korean Inst. Inf. Technol.*, vol. 20, no. 1, pp. 199–208, Jan. 2022.

[53] H. Xiong, S. Berkovsky, R. V. Sharan, S. Liu, and E. Coiera, "Robust vision-based workout analysis using diversified deep latent variable model," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2155–2158.

[54] V. Agarwal, K. Sharma, and A. K. Rajpoot, "AI based yoga trainer—Simplifying home yoga using mediapipe and video streaming," in *Proc. 3rd Int. Conf. Emerg. Technol. (INCET)*, May 2022, pp. 1–5.

[55] A. Chaudhari, O. Dalvi, O. Ramade, and D. Ambawade, "Yog-guru: Real-time yoga pose correction system using deep learning methods," in *Proc. Int. Conf. Commun. Inf. Comput. Technol. (ICCICT)*, Jun. 2021, pp. 1–6, doi: 10.1109/ICCICT50803.2021.9509937.

[56] Z. Wang, R. Liu, H. Zhao, S. Qiu, X. Shi, J. Wang, and J. Li, "Motion analysis of deadlift for trainers with different levels based on body sensor network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021, doi: 10.1109/TIM.2021.3062162.

[57] S. Oh and D. Kim, "Development of squat posture guidance system using kinect and wii balance board," *J. Inf. Commun. Converg. Eng.*, vol. 17, pp. 74–83, Jan. 2019.

[58] R. J. L. Virtucio and P. C. Naval, "Vision-based posture classification of the standard squat exercise using mediapipe pose," Ph.D. dissertation, Dept. Comput. Sci., Phillippines Diliman, Quezon City, Philippines, 2022.

[59] M. N. Hisham, M. F. A. Hassan, N. Ibrahim, and Z. M. Zin, "Mono camera-based human skeletal tracking for squat exercise abnormality detection using double exponential smoothing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 7, pp. 1–6, 2022, doi: 10.14569/IJACSA.2022.0130709.

[60] S. M. Remedios, D. P. Armstrong, R. B. Graham, and S. L. Fischer, "Exploring the application of pattern recognition and machine learning for identifying movement phenotypes during deep squat and hurdle step movements," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 364, Apr. 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00364

[61] N. Rungsawasdisap, A. Yimit, X. Lu, and Y. Hagihara, "Squat movement recognition using hidden Markov models," in *Proc. Int. Workshop Adv. Image Technol. (IWAIT)*, Chiang Mai, Thailand, 2018, pp. 1–4, doi: 10.1109/IWAIT.2018.8369648.

[62] N. Rungsawasdisap, X. Lu, A. Yimit, Z. Zhang, M. Mikami, and Y. Hagihara, "Squat movement recognition using convolutional neural network," in *Proc. SICE Annu. Conf.*, Hiroshima, Japan, Sep. 2019.

[63] M. Ota, H. Tateuchi, T. Hashiguchi, T. Kato, Y. Ogino, M. Yamagata, and N. Ichihashi, "Verification of reliability and validity of motion analysis systems during bilateral squat using human pose tracking algorithm," *Gait Posture*, vol. 80, pp. 62–67, Jul. 2020.

[64] S. Zhang, W. Chen, C. Chen, and Y. Liu, "Human deep squat detection method based on MediaPipe combined with YOLOv5 network," in *Proc. 41st Chin. Control Conf. (CCC)*, Jul. 2022, pp. 6404–6409.

[65] T. Yu and J.-H. Zou, "Automatic human gait imitation and recognition in 3D from monocular video with an uncalibrated camera," *Math. Problems Eng.*, vol. 2012, pp. 1–35, Jan. 2012.
[66] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021.
[67] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[68] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[69] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[70] J.-M. Frahm, K. Köser, and R. Koch, "Pose estimation for multi-camera systems," in *Pattern Recognition*. Tübingen, Germany: Springer, Aug./Sep. 2004, pp. 286–293.
[71] N. Eichler, H. Hel-Or, and I. Shimshoni, "Spatio-temporal calibration of multiple kinect cameras using 3D human pose," *Sensors*, vol. 22, no. 22, p. 8900, Nov. 2022.
[72] TemugeB. *Python-Stereo-Camera-Calibrate*. Accessed: May 25, 2023. [Online]. Available: https://github.com/TemugeB/python-stereo-camera-calibrate
[73] *BodyPose3D*. Accessed: May 25, 2023. [Online]. Available: https://github.com/TemugeB/bodypose3D

MUKUNDAN CHARIAR received the B.Tech. degree in mechatronics from the Manipal Institute of Technology, Manipal, Karnataka, India, in 2023. He is currently pursuing the master's degree in mechanical engineering with Carnegie Mellon University, Pittsburgh, PA, USA. He interned as a Trainee at Rex Engineering and Metal Industries, in 2022. He was the Vice-President of the Institution of Engineers Student Chapter, Manipal Institute of Technology. He is a member of the Institution of Engineers, India.

SHREYAS RAO received the B.Tech. degree in mechatronics from the Manipal Institute of Technology, Manipal, Karnataka, India, in 2023. He interned as a Trainee at Wagen Tunen Company, in 2022. Since 2023, he has been a Graduate Engineer Trainee at Royal Enfield, Chennai, Tamil Nadu, India.

ARYAN IRANI received the B.Tech. degree in mechatronics from the Manipal Institute of Technology, Manipal, Karnataka, India, in 2023. He interned as a Research and Development Intern at ACG Inspection Mumbai, in 2020. He was the Chief of Staff of a logistics company called Campus Express, in 2023.

SHILPA SURESH (Senior Member, IEEE) received the Ph.D. degree in computer vision from the National Institute of Technology, Karnataka, Suratkal, India, in 2018. She has authored more than 20 research articles. She has teaching experience of ten years and research experience of eight years. Her research interests include medical and satellite image processing, optimization algorithms, and machine learning techniques for various applications.

C S ASHA (Member, IEEE) received the Ph.D. degree in computer vision from the National Institute of Technology, Karnataka, Suratkal, India, in 2018. She has authored more than 20 research articles. She has teaching experience of 12 years and research experience of eight years. Her research interests include image processing in medical and satellite applications and computer vision for robotic applications. She is an Active Member of IET and ISTE.