Transformer-Integrated Pose Assessment Model with Real-Time Vocal Feedback for Autonomous Yoga Training

Dr.Swathi Agarwal
IT Department
CVR College of Engineering
Hyderabad – 50110, India
Email: swathiagarwal08@gmail.com

P Mohan Vamshi
IT Department
CVR College of Engineering
Hyderabad – 501510, India
Email: mohanvamshi2024@gmail.com

Abstract*—Yoga is a common practice to enhance the physical fitness, flexibility, balance and mental health. However, people who train without proper supervision tend to do an incorrect posture which can lead to injury and less effective training. Most existing systems for yoga pose recognition have considered pose classification and offered little assistance for real-time posture assessment and feedback. In this paper, a Transformer-Integrated Pose Assessment Model with Real-Time Vocal Feedback for Autonomous Yoga Training is proposed, which integrates Vision Transformer (ViT), MediaPipe Pose, joint-angle analysis and an intelligent feedback mechanism to achieve the automatic evaluation of yoga posture. The proposed system is composed of a standard webcam to capture the real-time video, a Vision Transformer to classify the Yoga pose and MediaPipe Pose to extract 33 body landmarks. The angles of the joints are calculated and compared with a reference posture database to assess the posture accuracy and produce a quantitative pose score. The system not only gives visual corrective feedback but also offers real-time oral feedback at opportune intervals so that the person can make corrections during their practice without a teacher's help. The proposed framework is tested using a popular and publicly available yoga pose dataset with 5 yoga poses and showed excellent classification accuracy (98.94% and 99.25% on training and validation set respectively), with reliable real-time inference

Keywords*—Keywords—Yoga Pose Assessment, Vision Transformer, MediaPipe Pose, Human Pose Estimation, Joint Angle Analysis, Real-Time Vocal Feedback.

I. Introduction

There are many benefits associated with yoga-flexibility, strength in muscles, posture, balance, reducing stress, and physical fitness are just a few. As yoga has become a popular fitness activity and an online education option, more and more people are practicing yoga without a professional teacher. But, yoga poses that are done incorrectly can lower the effectiveness of the exercise, and can make muscle strain, joint injuries and long-term postural problems more likely. Thus, real-time posture assessments and feedback on corrective actions are important research problem in the field of intelligent healthcare and computer vision applications.

In recent years, the ability of artificial intelligence, deep learning, and computer vision to accurately estimate human pose has paved the way for the creation of automated human pose estimation systems for fitness monitoring and rehabilitation. Although conventional convolutional neural network (CNN) based methods have shown to be effective in yoga pose classification, most of the existing systems have been capable of pose classification only from the predefined set of poses and have failed to assess the correctness of the posture and give real-time corrective feedback.

With the advent of Vision Transformers (ViTs), image classification has seen a great improvement, as they are able to capture long-range spatial dependencies in visual data in a self-attention manner. Likewise, MediaPipe Pose offers efficient real-time human pose estimation with a robust 33-pose body landmark extraction system that is very efficient in computation. The proposed approach combines image classification using the transformer and the analysis of the postures' skeletal landmarks, enabling the creation of an intelligent system that can recognize yoga poses and assess postures' quality in real-time.

In this paper, a Transformer-Integrated Pose Assessment Model with Real-Time Vocal Feedback for Autonomous Yoga Training is proposed, which integrates a Vision Transformer based on the classification of yoga poses and MediaPipe Pose based on landmark detection and estimation of yoga pose angles. The proposed framework will compare the angles formed between the user's joints with the pose database to determine a quantitative pose score and find out if the body is not aligned properly. Depending on the results of the evaluation, the system provides real-time visual guidance, as well as interval-based vocal feedback, so that the user can correct his/her posture during training without a constant instructor's supervision. The entire system is based on a common webcam so it can be used for autonomous yoga training at a low cost and is easily accessible.

After the classification of poses, the system uses MediaPipe Pose to perform a skeletonization of the anatomical body parts, obtaining the estimated location of thirty-three body points, which are corresponding the most significant joints, and key skeletal locations. These landmarks are used to give a lightweight yet robust representation of the practitioner's posture, without requiring any wearable sensors or specialised motion capture equipment. The coordinates of the landmarks extracted are then used to compute the joint angles associated with the significant body joint locations such as elbows and knees, which are used as quantitative descriptors of the alignment of the joint locations.

For checking the correctness of the pose, the calculated angles of these joints are checked with the database of reference poses having a set of ideal angle values for the supported Yoga poses. A pose score is calculated by analyzing the angular deviation between the detected pose and the reference pose, as well as a threshold value that has been set in advance.The angular deviation between the detected body pose and the reference pose, and a threshold value set in advance are used to calculate a pose score and determine which body joints are in need of correction. This evaluation strategy allows the proposed system to also evaluate the posture of the person performing the yoga pose, instead of only recognizing which pose he/she is performing, thereby allowing meaningful feedback on the quality of the person's pose.

An interactive feedback mechanism is integrated to enhance autonomous practice process user interaction. The evaluation results are used to generate visual corrective instructions on the display interface and the system also gives regular verbal feedback with pre-recorded audio messages. The visual and auditory feedback can help practitioners make immediate posture adjustments without constant guidance from a yoga expert, enhancing the accessibility and effectiveness of yoga training at home.

The modular structure of the proposed framework is another significant benefit, as it allows to enhance individual parts of the system like pose classification, landmark detection or posture evaluation independently without influencing the rest of the system. Additionally, the flexibility of this framework enables the easy addition of more yoga poses, more advanced evaluation techniques, or better deep learning techniques for future studies and possible deployment in the real world.

The key point of this work is the creation of a Transformer-Integrated Pose Assessment Model that efficiently recognizes yoga poses in real-time, evaluates their posture, and then provides corrective suggestions using just a regular RGB webcam. The proposed framework combines the use of a Vision Transformer to classify yoga poses and MediaPipe Pose to extract the 33 skeletal joints with high accuracy, and then calculates the joint angles and compares them with the reference posture database to quantify the correctness of the posture. Beyond pose recognition, the system is able to calculate a posture score, and recognize wrong body orientations, and provide personalized corrective feedback. This clever feedback system works on a visual as well as a verbal basis, and offers feedback on-the-go along with a periodic reminder of pre-recorded instructions, allowing the user to practice yoga without the need for constant supervision from a professional teacher. The proposed system can effectively support autonomous yoga training, home fitness, physiotherapy support, rehabilitation, and digital healthcare applications due to its light weight architecture, low hardware requirements, and real-time performance.

This paper continues with the following organization: Section II provides a summary of prior works related to Yoga pose recognition, Human pose estimation, and Deep learning based Posture Assessment techniques. Section III outlines the proposed methodology: preparing the datasets, pose classification using vision transformer, extracting landmarks from images using media pipe, computing angles between joints, evaluating the reference pose, and real-time vocal feedback mechanism. The results and discussion are given in Section IV. Finally, Section V summarizes the paper and discusses future research directions.

By combining the aforementioned pose classification with the MediaPipe skeletal landmark detection and joint-angle analysis, the proposed system can achieve more comprehensive posture evaluation than the traditional yoga pose recognition methods based on pose classification. Rather than just recognizing the executed yoga pose, the proposed approach continuously assesses body alignment, detects the incorrect joint positions, determines the posture correctness and gives individualized corrective feedback in the form of both visual and auditory cues. These features enhance the ease of use of standalone yoga training systems, allowing individuals to get guidance from an instructor without having to work with a real person in real time.

Beyond the conventional yoga training application, the proposed framework can be used to support many real-world applications, which can be related to the improvement of the posture assessment. The lightweight design, and low computation requirement, allows deployment on edge devices such as personal computers with only a standard RGB camera. This ability to access the framework from home makes it an ideal choice for physiotherapy support, the monitoring of rehabilitation progress, as well as smart health systems and fitness apps, where ongoing posture assessment and feedback can enhance user safety and training results.

II. LITERATURE REVIEW

This section presents a thorough review of the different Machine Learning, Deep Learning and Transformer Based methods reported in literature on human pose estimation, yoga pose recognition and posture assessment, explaining the methods, advantages, limitations and contributions in the development of intelligent yoga training systems.

One of the first comprehensive surveys of human pose estimation was given by Sigal et al. [1] where they discussed both model-based and learning-based methods for finding the locations of human body parts from images and videos. It talked about probabilistic graphical models and pictorial structures and spatial relationships between body parts and some of the problems involved such as body occlusion, viewpoint variations, and background clutter. This work laid the groundwork for the present pose estimation system, but its primary concern was hand-crafted representations and classical computing intensive and non-

generalizable optimization methods used to solve the problem in complex real world scenarios.

Yadav et al. [2] presented CNN-based real-time yoga recognition system for automatic yoga pose classification. It performed pre-processing on the captured image frames of the video sequences and extracted the spatial features from the image frames using deep convolutional layers, before conducting the classification of yoga postures. Experimental results showed that CNNs outperformed the traditional machine learning algorithms for yoga pose recognition by a large margin. However, the proposed framework mainly focused on pose classification and did not provide detailed posture deviation analysis or corrective feedback for improving user performance.

Rafi et al. [3] proposed an efficient CNN architecture for human pose estimation using the approach of producing heatmaps of the joints from the input images. Their model is competitive with other models on the benchmark datasets with relatively low complexity. The proposed framework successfully localized the body joints in a deep feature extraction and heatmap regression approach. However, the system focused mostly on joint localization, and higher-level posture assessment or interactive guidance systems for autonomous yoga training were not included.

Haque et al. [4] proposed a deep learning framework for exercise posture detection based on convolutional neural network (CNN). The proposed system has extracted discriminative features from the exercise images and classified various fitness activities with the improved recognition accuracy. The framework demonstrated better performance than conventional machine learning approaches in identifying exercise postures. However, it primarily acknowledged exercise categories but did not judge whether the exercise posture of each joint was correct or not and did not provide corrective assistance during exercise.

Islam et al. [5] presented a yoga posture recognition system based on Microsoft kinect to get the 3D skeletal joint information to analyze yoga in real-time. The accuracy of pose recognition was improved by using the depth sensors, which helped in localization of body joints, as compared to the RGB-image-based approach. While the skeletons could be tracked successfully with the Kinect-based systems, the systems were expensive, and the use of the skeletons was not feasible for home-based yoga training or mobile healthcare applications.

Patil et al. [6] proposed the Yoga Tutor framework which employed SURF techniques to visualize and analyze yoga poses from a feature extraction point of view. The proposed approach offered a simple self-learning platform for yoga practitioners, and illustrated the potential of computer vision techniques for posture analysis. However, this handcrafted feature extraction strategy was very sensitive to changes in lighting, camera viewpoints, and complex background.

To give a comprehensive overview of the progress in human pose estimation from monocular images, Gong et al. [7] conducted a survey of the major techniques developed for human body joint estimation from monocular images, using computer vision algorithms. We discussed top-down/ bottom up body landmark detection and classified pose estimation methods into traditional model-based and modern deep learning-based methods. The authors studied various benchmark datasets, evaluation metrics, and problems such as body occlusion, illumination variation, variation of viewing angles, and complex backgrounds. Additionally, the authors pointed out the development of CNN towards pose estimation precision and efficiency. While the survey gives a thorough overview of the current methodologies and future directions for research, it does not offer a novel pose estimation framework. Also, the transformer-based architectures and real-time corrective feedback mechanisms were not used, and there was scope for more intelligent and interactive pose assessment systems.

Ning et al. [8] put forward an articulated human pose estimation and tracking framework from the top down, by first detecting the human subject by an object detection model and then in turn estimating the location of the body joints. The proposed approach shows that localization accuracy is improved by concentrating on each human being in the individual and the ambiguity is reduced, if two or more people are detected in the same image. This framework incorporates human detection and pose estimation to obtain good tracking performance in complicated scenes and to enhance the performance of pose estimation in general. Experimental results showed that the proposed technique outperformed some of the traditional bottom-up techniques in terms of accuracy on the benchmark datasets. The approach does need an accurate human detector as a preprocessing stage, however, which adds computational complexity and reduces the efficiency of the approach for real-time applications. Furthermore, the framework is mainly used for joint localization, and it didn't consider the correctness of the posture and didn't propose the corrective posture for fitness or yoga training applications.

Gupta et al. [9] proposed a human body pose estimation system that estimates the location of the human body joints in a sequence of images using computer vision methods. The proposed system is capable of identification of various anatomical landmarks and reconstruction of human skeletal structure for motion analysis and activity recognition. The main objective of the framework is to enhance the localization, in different environmental conditions, of human body joints in order to facilitate applications like surveillance, gesture recognition and human-computer interaction. The proposed approach is mainly focused on the detection of the body joints and not much on the quality of the posture in terms of semantics. Additionally, it is lacking in the integration of deep learning based spatial feature learning, and does not include intelligent posture assessment and interactive correction for autonomous yoga practice.

Agarwal and Triggs [10] proposed a human pose estimation approach in 3D, which was silhouette based feature extraction combined with Relevance Vector Regression (RVR). The proposed framework has the ability to estimate locations of body joints in 3D from 2D silhouette images without using explicit kinematic body models. Through learning the relationship between image features and body joint positions, the method shown the pose reconstruction accuracy was more than the conventional model-based method. The learning-based regression method also alleviated the need for handcrafted geometric models and demonstrated promising results for 3D human pose estimation. The proposed method is however sensitive to background clutter, illumination variations and the appearance of clothes and would require extraction of accurate silhouette. Moreover, it lacks features such as deep learning, transformer-based feature learning, real-time skeletal landmark extraction, and intelligent posture analysis and feedback that are crucial to the development of autonomous yoga training applications.

To achieve efficient body landmark detection on resource constrained devices, Bazarevsky et al. [11] introduced the lightweight real-time human pose estimation system BlazePose. The framework is able to accurately estimate 33 anatomical body landmarks from RGB images using a single camera, with low computational complexity and high inference speed. BlazePose's real-time performance was remarkable for different human actions and it's skeleton tracking was reliable for mobile and desktop applications. Although the framework is effective for landmark detection, it is a basic framework that mainly concentrates on body keypoint estimation and fails to assess the correctness of posture, calculate deviations of joint angles and generate intelligent corrective feedback. Furthermore, BlazePose lacks the capability of pose classification using a transformer, along with autonomous vocal guidance, which reduces its ability to assess complete postures in yoga.

To tackle this, Xu et al. [12] introduced a transformer-based human pose estimation framework named ViTPose, which is an extension of Vision Transformer architectures for generic body pose estimation. The framework uses transformer-based feature extraction, which is able to encode both local and global spatial information, leading to better localization accuracy of human body joints across several benchmark datasets. ViTPose++ used efficient architectural improvements and optimized training methods to reach the state-of-the-art performance while retaining good generalization capabilities under various human poses and imaging conditions. Though the proposed framework is a big leap in developing pose estimation using transformer, it's primary goal is landmark localization rather than quality assessment of the pose. Moreover, it doesn't feature any of the essentials of an independent yoga training system, like reference pose comparison, joint-angle assessment, pose score, or real-time visual and vocal corrective feedback.

III. PROPOSED METHODOLOGY

The proposed system aims to create an autonomous yoga pose assessment system that integrates computer vision, transformer-based deep learning, human pose estimation, joint-angle analysis, and intelligent vocal feedback. The framework works on a regular RGB webcam and doesn't need any kind of wearables or unique motion capturing gear. First, a Vision Transformer (ViT) is used to categorize the selected yoga pose from the video frames taken from the video. Then, MediaPipe Pose identifies 33 key joint points on the body, which are used to calculate some of the key angles for the elbows and knees. These joint angles are then compared to a reference posture database to check the correctness of one's posture and to determine a quantitative pose score. The system is able to detect posture deviations and provide real-time visual corrective guidance and periodically provide prerecorded vocal guidance to correct posture during autonomous yoga exercise. The complete process of the suggested framework are depicted in Fig. 1.

3.1 System Architecture Overview

The proposed system architecture includes nine main modules that contribute to the overall system in carrying out real time classification of yoga poses, real-time assessment of yoga poses and intelligent corrective feedback. Using OpenCV, the pictures of the video are captured in continuous stream from a standard RGB web cam as frames. The captured frame is resized and passed to trained Vision Transformer model, which categorizes the performed yoga pose into the yoga category from the pre-defined list. At the same time, MediaPipe Pose estimates the position of 33 body joints of the user. These landmarks are used to determine key joint angles and the angle data is compared to a database of expected joint angles for each yoga posture. These deviations from the standard are compared, a quantitative score of "pose" is determined, and incorrectly aligned body joints are identified. Lastly the system offers visual corrective feedback on the screen and also offers voice feedback periodically while practicing so the individual can correct his/her posture. The entire layout is shown in Fig. 1.

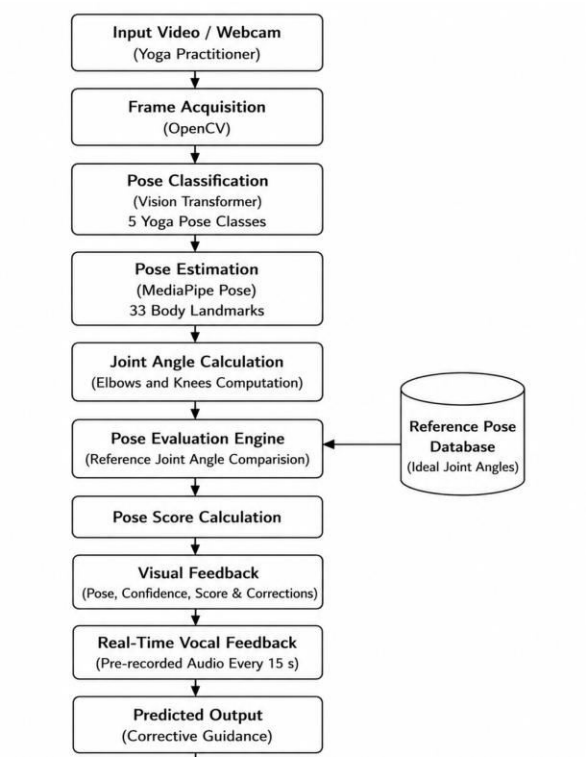

Fig. 1. Overall system architecture.

3.2 Video Acquisition Component

The Video Acquisition component is used to continuously capture real-time RGB frames using a standard webcam via OpenCV library. All frames captured are resized and then converted to the appropriate format prior to being fed into the Vision Transformer classification model and MediaPipe Pose estimation framework. Continuous frame acquisition allows smooth real-time processing while ensuring a suitable image quality for accurate pose recognition and landmark extraction. The system requires only a conventional RGB camera, which is capable of being easily installed and is low cost in comparison, so that it can be used as a low-cost and easily deployable solution for autonomous yoga training.

3.3 Vision Transformer Pose Classification Component

The main model used for classification of yoga poses is the Vision Transformer (ViT). The captured RGB frames are segmented into fixed-size image patches which are encoded into feature vectors and then passed through a series of transformer encoder layers with a self-attention mechanism. The Vision Transformer processes the spatial relationships between all image data, unlike traditional convolutional neural networks, which makes it better at identifying similar yoga poses from a visual perspective. The trained model classifies the one of the predefined yoga pose classes along with its confidence score which is then passed to the posture evaluation module.

3.4 MediaPipe Pose Landmark Detection

MediaPipe Pose classifies the pose and then makes real-time predictions for 33 anatomical body landmarks of the identified practitioner. The following are the significant joints of the bones: Shoulders, elbows, wrists, hips, knees, ankles, feet. MediaPipe's lightweight architecture allows for accurate landmark detection with low computational complexity and real-time performance. The obtained landmarks coordinates are used as a basis for further calculation of the joint angles and a posture evaluation.

3.5 Joint Angle Computation

The Joint Angle Computation component outputs the important angles between various body joints detected by MediaPipe Pose with coordinates of skeleton landmarks. Posture evaluation is done on four key anatomical joints – left elbow, right elbow, left knee, right knee. The angles between the joints are calculated geometrically, using vectors, in degrees. These angles quantify the body alignment and are eventually used to assess the body posture.

3.6 Reference Pose Evaluation Engine

The Reference Pose Evaluation Engine compares computed joint angles with a reference pose database pre-defined with ideal joint angles for the supported yoga poses. Each joint is checked to see if it is properly aligned to a predefined angular tolerance. The results of the comparison are used to calculate a quantitative score of pose and to determine which body parts need to be corrected. This evaluation process allows the framework to evaluate posture quality instead of just identifying the executed yoga pose.

3.7 Visual Feedback Component

The Posture Evaluation results are used as input to the Visual Feedback component, which displays the predicted yoga pose, classification confidence, posture score, and posture recommendations on the UI. Descriptive messages are used to indicate body joints that need to be corrected, which helps practitioners become aware of deviations in posture when practicing yoga. The continuous visual guidance allows the user to implement posture corrections while practicing.

3.8 Real-Time Vocal Feedbac

Real-time Vocal Feedback delivers intelligent audio feedback from the results of posture evaluation. Rather than using a Text-to-Speech engine for dynamic speech generation, the proposed system has a set of prerecorded audio messages, each of which is associated with a commonly used posture correction, e.g., making the elbows straight or knees straight, and fixing the body alignment. The system checks the current posture every 15 seconds, and if necessary, plays the highest priority corrective instruction every 15 seconds. A positive sound message is played when the posture is done right, promoting receiving correct posture. Periodic vocal feedback and visual guidance provide an interactive, independent yoga training experience, but at the

same time, decreases user dependency on professional trainers.

IV. Results and Discussion

The proposed Transformer-Integrated Pose Assessment Model with Real-Time Vocal Feedback was tested using the Kaggle Yoga Pose dataset of five classes of yoga poses, DownDog, Goddess, Plank, Tree and Warrior II. The Vision Transformer model was trained for 20 epochs and then combined with the MediaPipe Pose library for extracting skeletal landmarks and evaluating posture in real time using the angles of the joints. Training and validation accuracy, training and validation loss, confusion matrix, per-class accuracy, and standard classification metrics such as Accuracy, Precision, Recall and F1-score were used to evaluate the overall performance of the proposed framework. In addition, the proposed framework was qualitatively assessed with respect to correctly classified and misclassified yoga pose samples to determine the robustness of the proposed framework with various body poses and viewing conditions.

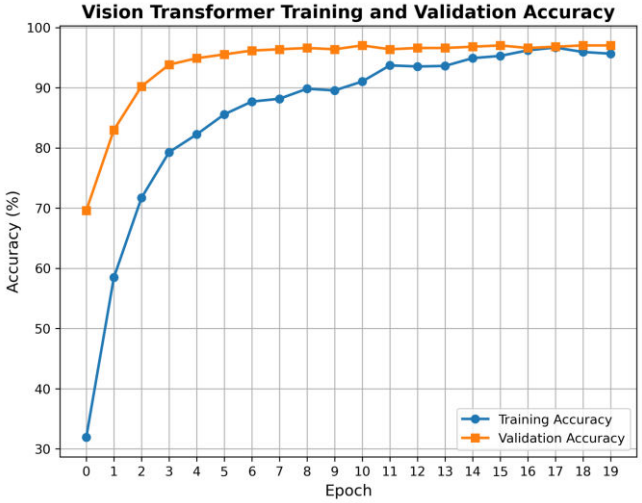

Fig. 2. Vision Transformer Training and Validation Accuracy.

The accuracy of the proposed Vision Transformer also shows learning performance during the training and validation process, as shown in Fig. 2. The training accuracy gradually rose from around 32% for the first epoch to over 96% for the last epoch, which shows that the model has learned these features from the yoga pose images. Likewise, the validation accuracy rose quickly to almost 70% and reached above 97% after a few epochs. The similarity between the training curve and the validation curve suggests that the proposed model exhibited good generalization to unseen data, with minimal overfitting. Both the curves are indeed smooth, which reflects the ability of the Vision Transformer to learn global spatial relationships necessary for precise yoga pose classification.

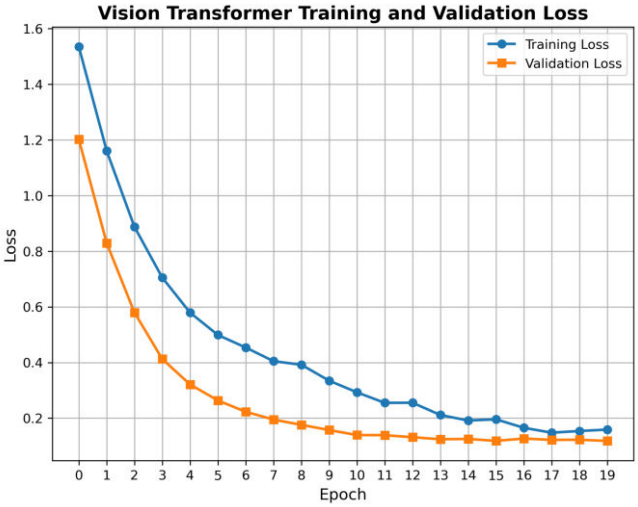

Fig. 3. Vision Transformer Training and Validation Loss.

Figure 3 shows the trend of training and validation loss over the course of learning. Both loss curves continuously decreased with increasing epochs, indicating stable optimization of the proposed model. The training loss improved steadily from ~1.53 to below 0.16, and the validation loss improved from ~1.20 to almost 0.12. The lack of noticeable differences or oscillations between the two lines suggests that the Vision Transformer converged smoothly and performed well with little divergence. The results of these observations confirm that the proposed framework was able to reduce the classification errors in the training process.

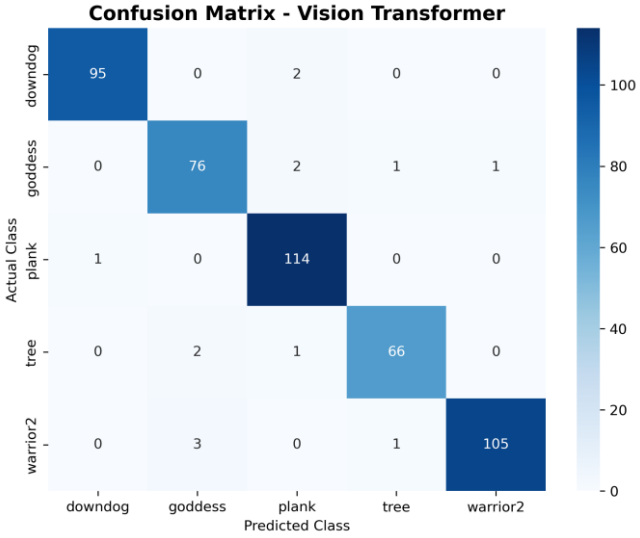

Fig. 4. Confusion Matrix of the Vision Transformer Model.

The confusion matrix in Fig. 4 presents the classification accuracy of the proposed Vision Transformer for the five yoga classes. The majority of test samples are correctly classified, thus the dominant diagonal elements of the matrix are present. The confusion between poses which are visually similar, such as Goddess and Warrior II, or Tree and Goddess, is very limited. The recognition rates of DownDog and Plank were very high with very few misclassifications. In general, the proposed model has low

classification error and accurately classifies yoga postures, as shown in the confusion matrix.

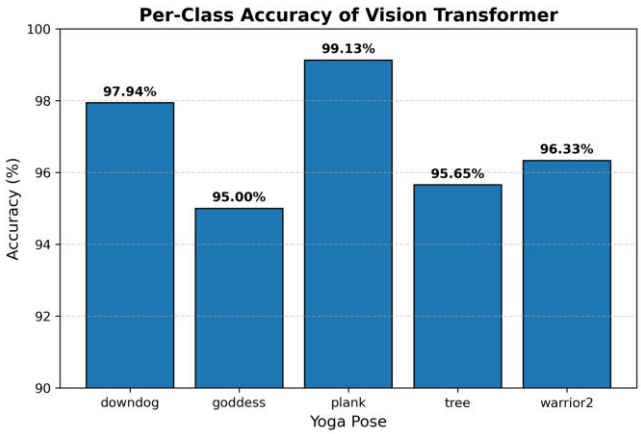

Fig. 5. Per-Class Classification Accuracy.

Figure 5 shows the classification accuracy and results for the different classes of yoga poses. Of these poses, Plank got the highest classification accuracy (99.13%), DownDog followed with 97.94%, Warrior II with 96.33%, Tree with 95.65% and Goddess with 95.00%. The average error for both five pose categories is very low, indicating the ability of the Vision Transformer model to accurately extract the discriminative spatial features of similar yoga poses. The results in this section also showed that the present scheme is balanced for various body configurations.

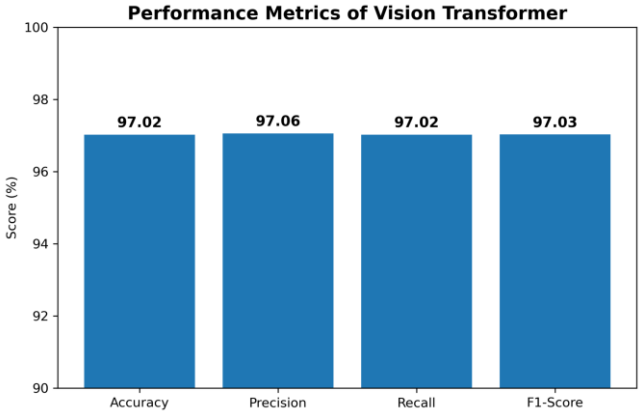

Fig. 6. Overall Performance Metrics.

A summary of the quantitative performance of the proposed framework is shown in Fig. 6. The classification accuracy, Precision, Recall, and F1 score obtained by the proposed framework were 97.02%, 97.06%, 97.02%, and 97.03%, respectively, for the Vision Transformer classifier. These findings show that there is an equal classification rate between false-positive and false-negative rates that are very low. The good evaluation scores shows that the transformer based architecture is able to effectively capture the global spatial relationships of body joints, thereby achieving robust pose recognition of yoga poses and accurate posture assessment.

V. CONCLUSION

This study proposes a Transformer-Integrated Pose Assessment Model with Real-Time Vocal Feedback for Autonomous Yoga Training as an intelligent paradigm for real-time yoga pose recognition, posture assessment, and guidance to correct postures. The system proposed here combines a Vision Transformer (ViT) network for yoga pose classification, MediaPipe Pose for real-time extraction of 33-body skeleton points, computation of the joints angles, a reference pose evaluation engine, and an interval-based vocal feedback system to create an effective and self-governing yoga training system. The proposed framework utilizes both transformer-based global feature learning and skeletal landmark analyses to enhance the recognition accuracy of yoga poses and assess the quality of the posture and misalignment of the body. The miniaturization of the webcam-based system dispenses with the need to wear (wearable) sensors or use special motion capture equipment, enabling cost-effective and easy-to-reach autonomous yoga practice.

The proposed framework was tested for its performance with several experimental metrics. The proposed framework has an overall classification accuracy of 97.02% with Precision of 97.06%, Recall of 97.02%, and F1-Score of 97.03% on five Yoga Poses Classes. The training and validation accuracy and loss curves showed satisfactory convergence without any signs of overfitting, and the confusion matrix and per-class accuracy results proved that the model is robust in distinguishing similar yoga poses. In addition, the joint-angle based pose evaluation engine combined with the periodic prerecorded voice feedback, allowed real-time posture correction which improved the usability of the framework in autonomous yoga training.

Finally, the proposed Transformer-Integrated Pose Assessment Model with Real-Time Vocal Feedback is shown to be effective in combining the transformer-based deep learning approach with skeleton landmark analysis for intelligent assessment of yoga posture. This framework integrates pose classification, posture scoring, visual corrective feedback, and interval-based voice feedback into a single system to offer a correct, reliable and user-friendly answer for autonomous yoga practise. The framework can be applied to digital healthcare, fitness training, physiotherapy rehabilitation and personalized wellness platforms with great potential. The framework will be further extended to accommodate more yoga poses, dynamic pose sequence analysis, adaptive feedback for every participant, mobile deployment and real-time monitoring of the yoga positions through the cloud.

References

[1] L. Sigal, "Human Pose Estimation," Encyclopedia of Computer Vision, Springer, 2011.
[2] S. Yadav, A. Singh, A. Gupta, and J. Raheja, "Real-Time Yoga Recognition Using Deep Learning," Neural Computing and Applications, vol. 31, no. 12, pp. 9349–9361, 2019.

[3] U. Rafi, B. Leibe, J. Gall, and I. Kostrikov, "An Efficient Convolutional Network for Human Pose Estimation," Proceedings of the British Machine Vision Conference (BMVC), 2016.
[4] S. Haque, A. Rabby, M. Laboni, N. Neehal, and S. Hossain, "ExNET: Deep Neural Network for Exercise Pose Detection," Recent Trends in Image Processing and Pattern Recognition, Springer, pp. 186–197, 2019.
[5] M. Islam, H. Mahmud, F. Ashraf, I. Hossain, and M. Hasan, "Yoga Posture Recognition by Detecting Human Joint Points in Real Time Using Microsoft Kinect," IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017.
[6] S. Patil, A. Pawar, and A. Peshave, "Yoga Tutor: Visualization and Analysis Using SURF Algorithm," IEEE Control and System Graduate Research Colloquium, pp. 43–46, 2011.
[7] W. Gong, X. Zhang, J. González, A. Sobral, T. Bouwmans, C. Tu, and H. Zahzah, "Human Pose Estimation from Monocular Images: A Comprehensive Survey," Sensors, vol. 16, no. 12, Art. no. 1966, 2016.
[8] G. Ning, P. Liu, X. Fan, and C. Zhan, "A Top-Down Approach to Articulated Human Pose Estimation and Tracking," European Conference on Computer Vision (ECCV) Workshops, 2018.
[9] A. Gupta, T. Chen, F. Chen, and D. Kimber, "Systems and Methods for Human Body Pose Estimation," U.S. Patent 7,925,081 B2, Apr. 12, 2011.
[10] A. Agarwal and B. Triggs, "3D Human Pose from Silhouettes by Relevance Vector Regression," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 882–888, 2004.
[11] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-Time Body Pose Tracking," arXiv preprint arXiv:2006.10204, 2020.
[12] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose++: Vision Transformer for Generic Body Pose Estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 2, pp. 1212–1230, Feb. 2024, doi: 10.1109/TPAMI.2023.3330016.