Computer vision-based foot contact detection for long jump using a monocular normal-speed camera

Yangtao Fang (IP Paris), Qi Gan (Telecom Paris, IP Paris), Sao Mai Nguyen (Ensta, IP Paris)

1 Introduction

Our research analyzes the long jump, particularly the approach run, which has two phases: acceleration and zeroing-in[1]. Approach drills enhances long jump performance by improving approach speed, explosive power, and neuromuscular coordination. A key part of training is ensuring proper foot contact, which affects stride efficiency and injury risk. However, without high-speed cameras, evaluating foot-ground contact using traditional observation methods is imprecise and time-consuming. Advances in deep learning, particularly Vision Transformer (ViT), which splits each image into a sequence of tokens to model global relations for classification[2]. This study proposes a hybrid ViT-LSTM model to detect the degree of foot-ground contact in long jump athletes using images captured by a monocular normal-speed camera, achieving an accuracy of 91.87% and 8.18 milliseconds/frame processing speed with low computational resources (321 TOPS).

2 Methodology

2.1 Dataset

The dataset consists of 30 video clips[3], each exported as a frame image sequence, totaling approximately 3,000 images. The dataset is captured using a monocular normal-speed camera at only 25 frames per second, leading to motion blur that obscures athletes' movements, especially rapid leg motions. The training and validation sets include 25 video clips. The test set consists of 5 video clips. 2D skeleton data was extracted using ViTPose[4] and manually corrected. The distribution of the 19 joint points is shown in Figure 1.

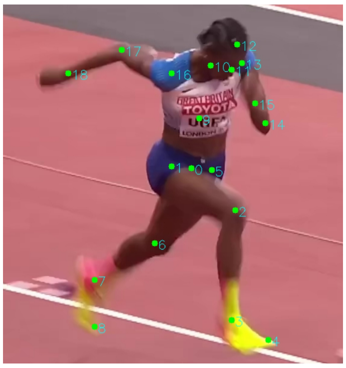


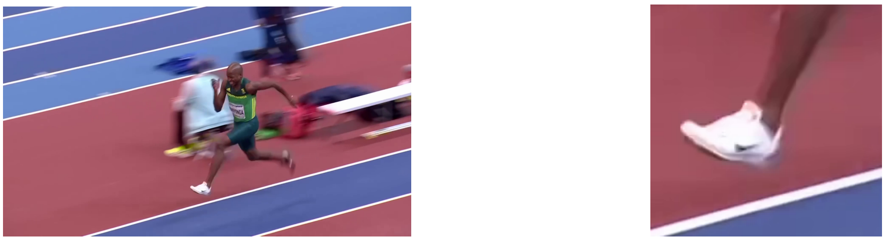
Figure 1: Joint points

2.1.1 Pose Normalization

To enhance the robustness of the pose data and reduce variability due to absolute positioning, we normalized the 2D joint coordinates by setting joint 0 (the hip) as the coordinate origin for each frame. This normalization process subtracts the hip coordinates from all other joint coordinates within the same frame, effectively re-centering the pose data.

2.1.2 Image processing

To address noise and distractions in full images, such as the presence of other individuals, we identified the landing foot by comparing the y-coordinates of the left and right foot joints (indexes 3 and 7). The joint with the larger y-coordinate, indicating a lower position in the image, was selected. A 224×224 pixel image was then cropped, centered on the selected joint. The resulting ankle image is shown in Figure 2b.



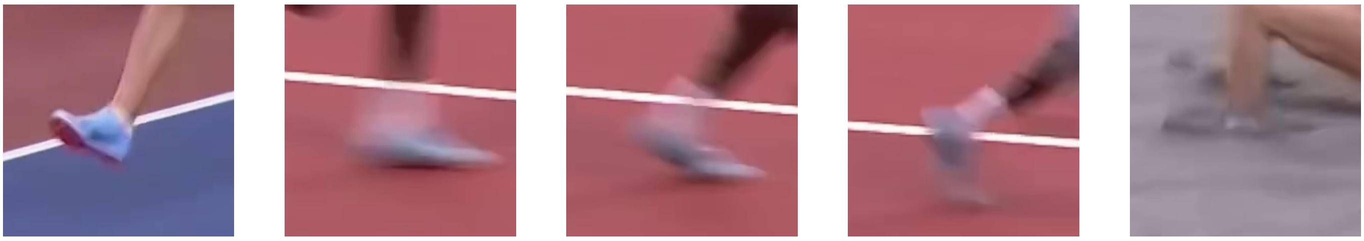
(a) Full image

(b) Cropped image

Figure 2: Long Jump Image Dataset

2.1.3 Label Definition

Existing studies often limit foot contact detection to a binary classification (contact vs. no contact), lacking granularity in analyzing contact degrees. To enable detailed analysis of athletes' movement techniques for coaching, we defined five contact labels: 0 (no ground contact), 1–3 (progressive stages of ground contact, from initial touch to just before lift-off), and 4 (contact with the sandpit). Figure 3 illustrates examples of these contact stages.



(a) Label 0 (foot not in contact with the ground)

(b) Label 1 (soles of feet touching ground)

(c) Label 2 (heel lifted, keep the forefoot on the ground)

(d) Label 3 (toes in contact with the ground)

(e) Label 4 (foot in contact with the sandpit)

Figure 3: Label Definition

2.1.4 Data preprocessing

To increase data diversity and prevent rapid overfitting on this small dataset, the training set employed data augmentation, which includes resizing to 224 × 224, color jittering, random small-angle rotation, and normalization with ImageNet statistics[5]. The validation set was only resized and normalized. We pads sequences of varying lengths to the maximum length in a batch, enabling efficient batch processing with a padding label of -100 for ignored indices.

2.2 Model Architecture

The proposed hybrid ViT-LSTM model integrates a Vision Transformer for spatial feature extraction, an attention-based fusion mechanism for combining image and pose features[6], and a bidirectional LSTM for temporal modeling[7]. The model architecture is shown in Figure 4.

The ViT is initialized with pre-trained weights to speed up convergence and reduce computational resource consumption. The pretrained ViT backbone extracts 768-dimensional spatial features from cropped ankle images, while pose data (flattened to 38 dimensions from 19 joint points) is projected to 768 dimensions. The Attention Fusion module concatenates the two feature sets and computes attention weights through a sequence of linear layers with Tanh and Sigmoid activations. These weights enable a weighted combination of the feature sets, producing fused features that retain a 768-dimensional feature size. This approach leverages learned attention weights to dynamically balance the contributions of image and pose information, followed by a residual connection with the ViT features and layer normalization to stabilize the output.

The fused features are processed by a bidirectional LSTM, followed by a classifier predicting five foot contact states per frame. This bidirectional approach ensures that the state of foot contact in one frame is informed by both preceding and following frames, capturing the full context of the movement.

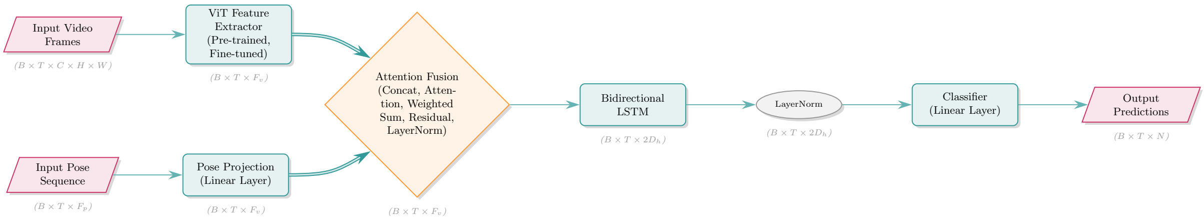

Input Video Frames
$(B \times T \times C \times H \times W)$
ViT Feature Extractor (Pre-trained, Fine-tuned)
$(B \times T \times F_v)$
Input Pose Sequence
$(B \times T \times F_p)$
Pose Projection (Linear Layer)
$(B \times T \times F_v)$
Attention Fusion (Concat, Attention, Weighted Sum, Residual, LayerNorm)
$(B \times T \times F_v)$
Bidirectional LSTM
$(B \times T \times 2D_h)$
LayerNorm
$(B \times T \times 2D_h)$
Classifier (Linear Layer)
Output Predictions
$(B \times T \times N)$

Figure 4: Hybrid ViT-LSTM model architecture. *Notation:* B: batch size, T: sequence length, C: number of channels, H: frame height, W: frame width, F_p: pose feature dimension, F_v: ViT feature dimension, D_h: hidden size (per direction), N: output classes.

2.3 Training Setup

We freeze most ViT parameters, unfreezing only the last two transformer blocks and normalization layer for fine-tuning, thereby reducing trainable parameters and lowering computational resource requirements. The model is trained using 5-fold cross-validation. Each fold is trained on 20 videos and validated on 5 videos. By rotating the training and validation sets across all folds, the model is trained and evaluated on the entire dataset, maximizing the use of limited data and providing a more robust estimate of its generalization performance[8]. The AdamW optimizer is employed with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}, paired with a ReduceLROnPlateau scheduler (factor=0.5, patience=5) to adjust the learning rate based on validation loss. To tackle the class imbalance where airborne frames outnumber foot-contact frames, we employ a weighted cross-entropy loss, where class weights are automatically computed based on the training data distribution[9].

3 Comparison experiments

We tested seven different model architectures on the GeForce RTX 4070 Laptop GPU (8GB memory, 321 TOPS). Our proposed model architecture (model e) achieves the highest classification accuracy with a processing speed of 8.18 milliseconds per frame. For comparison, we trained a model f without an attention mechanism but with the same configuration, and got an average validation accuracy of 0.895. This is significantly lower than the model e with the attention mechanism. The results of the comparison experiments are shown in Table 1.

Table 1: Comparison of Experimental Results for Foot Contact Detection

Model Architecture	**Input**	**Feature**	**Pose Norm**	**Fusion Method**	**Epochs**	**Val. Accuracy**	**Figures**
ViT (a)	Full Image	Image	No	None (768D)	20	0.8433	5, 6, 7
ViT + LSTM (b)	Full Image	Image	No	None (768D)	30	0.8604	8, 9, 10
ViT + LSTM (c)	Cropped Image	Image and Pose	No	Addition (768D)	100	0.8625	11, 12
ViT + LSTM (d)	Cropped Image	Image and Pose	No	Concatenation (1536D)	100	0.9083	13, 14, 15
ViT + LSTM (e)	Cropped Image	Image and Pose	Yes	Attention (768D)	40 (5-fold)	0.9187	16, 17, 18
ViT + LSTM (f)	Cropped Image	Image and Pose	Yes	Concatenation (1536D)	40 (5-fold)	0.8950	19, 20
ViT + Transformer (g)	Cropped Image	Image and Pose	Yes	Attention (768D)	40 (5-fold)	0.8783	21, 22, 23

4 Conclusion

Experimental results show that techniques such as cropping key parts of images, pose normalization, cross-validation, feature fusion, and attention mechanisms can help improve the model's classification accuracy of foot contact. The proposed hybrid ViT-LSTM model (model e) maintains high detection accuracy (91.87%) and fast processing speed (8.18 milliseconds per frame) with low computational resources even under low-frame-rate video conditions, opening up new possibilities for the deployment of deep learning-based automated motion analysis in resource-constrained scenarios.

References

[1] Hubert Makaruk, Jared Porter, Marcin Starzak, and Edward Szymczak. An examination of approach run kinematics in track and field jumping events. *Polish Journal of Sport and Tourism*, 23:82–87, 06 2016.

[2] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.

[3] Qi Gan, Sao Mai Nguyen, Mounîm A. El-Yacoubi, Eric Fenaux, and Stéphan Clémençon. Human pose estimation based biomechanical feature extraction for long jumps. In *2024 16th International Conference on Human System Interaction (HSI)*, pages 1–6, 2024.

[4] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation, 2022.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R. Hershey, and Tim K. Marks. Attention-based multimodal fusion for video description, 2017.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[8] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[9] Trong Huy Phan and Kazuma Yamamoto. Resolving class imbalance in object detection with weighted cross entropy losses, 2020.

A Appendix

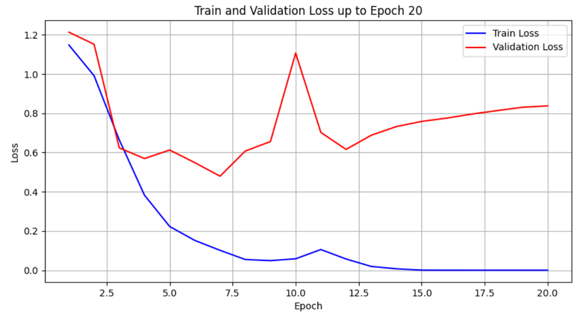


Figure 5: Training and Validation Loss over 20 Epochs (ViT with Full Images Using Only Image Features)

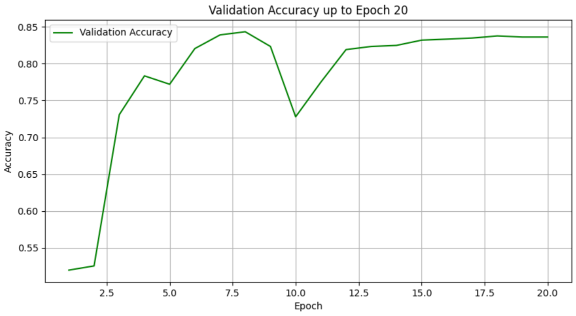


Figure 6: Validation Accuracy over 20 Epochs (ViT with Full Images Using Only Image Features)

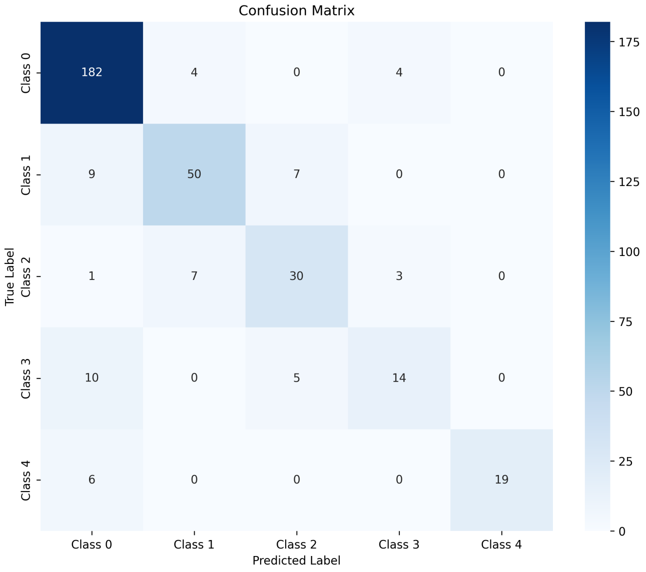


Figure 7: Confusion matrix (ViT with Full Images Using Only Image Features)

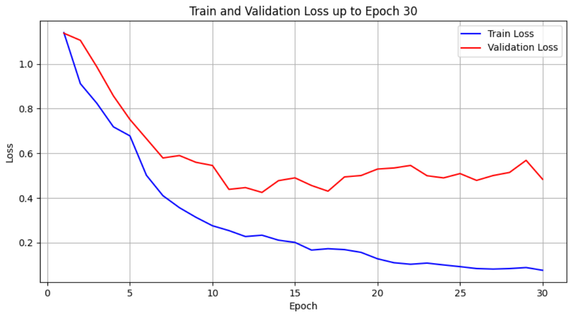


Figure 8: Training and Validation Loss over 30 Epochs (ViT and LSTM with Full Images Using Only Image Features)

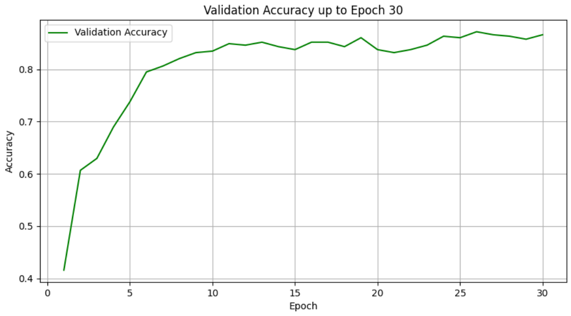


Figure 9: Validation Accuracy over 30 Epochs (ViT and LSTM with Full Images Using Only Image Features)

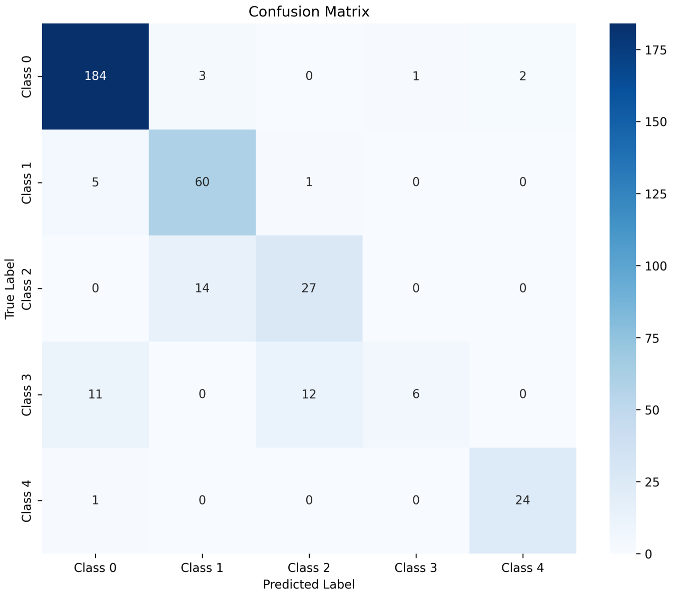


Figure 10: Confusion matrix (ViT and LSTM with Full Images Using Only Image Features)

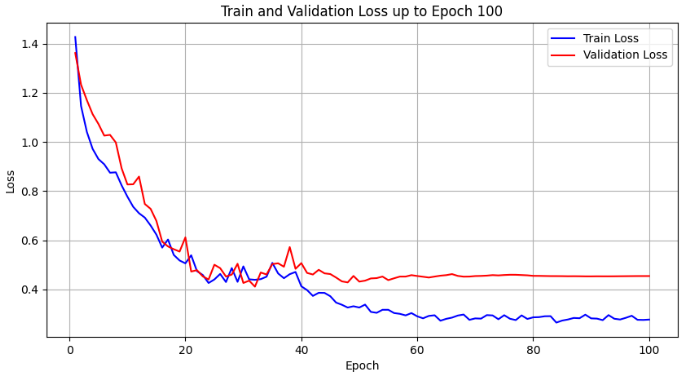


Figure 11: Training and Validation Loss over 100 Epochs (ViT and LSTM with Cropped Ankle Images Using Simple Addition of Joint and Image Features)

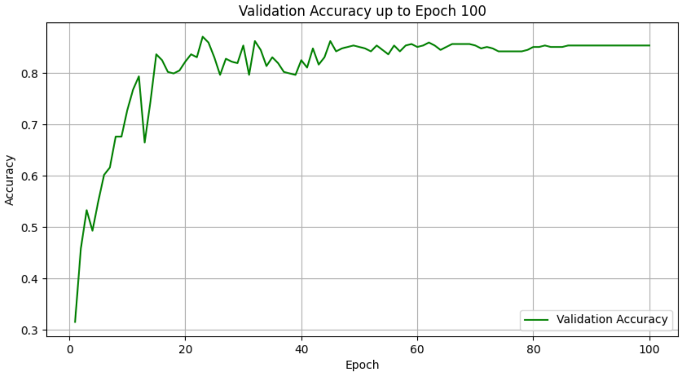


Figure 12: Validation Accuracy over 100 Epochs (ViT and LSTM with Cropped Ankle Images Using Simple Addition of Joint and Image Features)

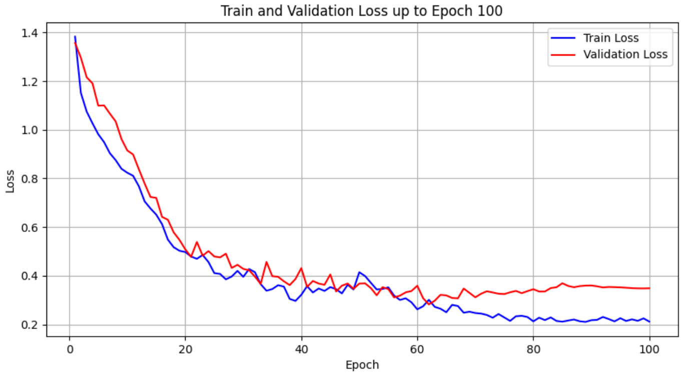


Figure 13: Training and Validation Loss over 100 Epochs (ViT and LSTM with Cropped Ankle Images Using Concatenation of Joint and Image Features)

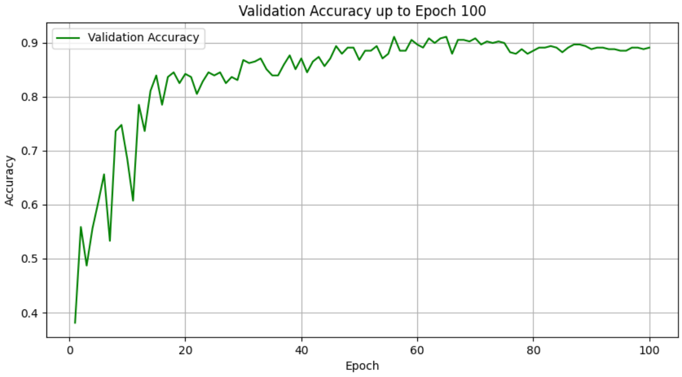
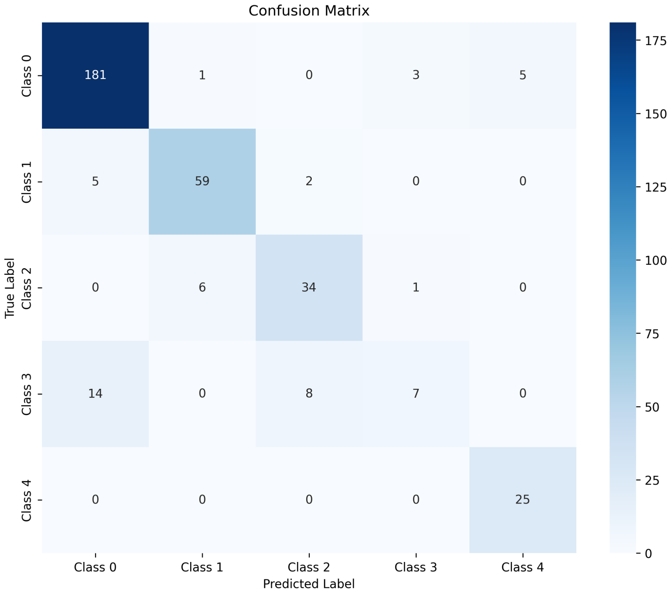
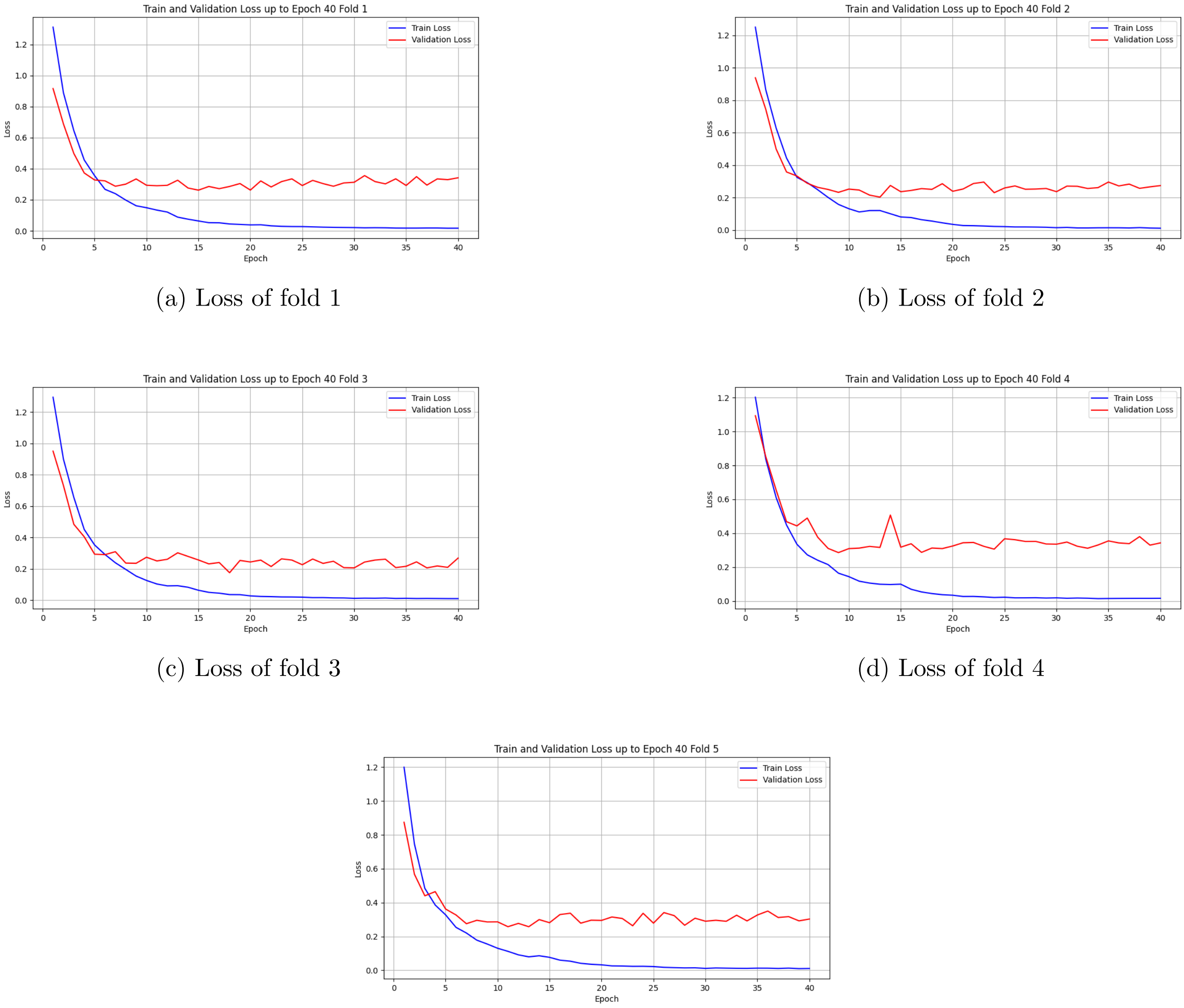


Figure 14: Validation Accuracy over 100 Epochs (ViT and LSTM with Cropped Ankle Images Using Concatenation of Joint and Image Features)



	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	181	1	0	3	5
Class 1	5	59	2	0	0
Class 2	0	6	34	1	0
Class 3	14	0	8	7	0
Class 4	0	0	0	0	25

Figure 15: Confusion matrix (ViT and LSTM with Cropped Ankle Images Using Concatenation of Joint and Image Features)



(a) Loss of fold 1

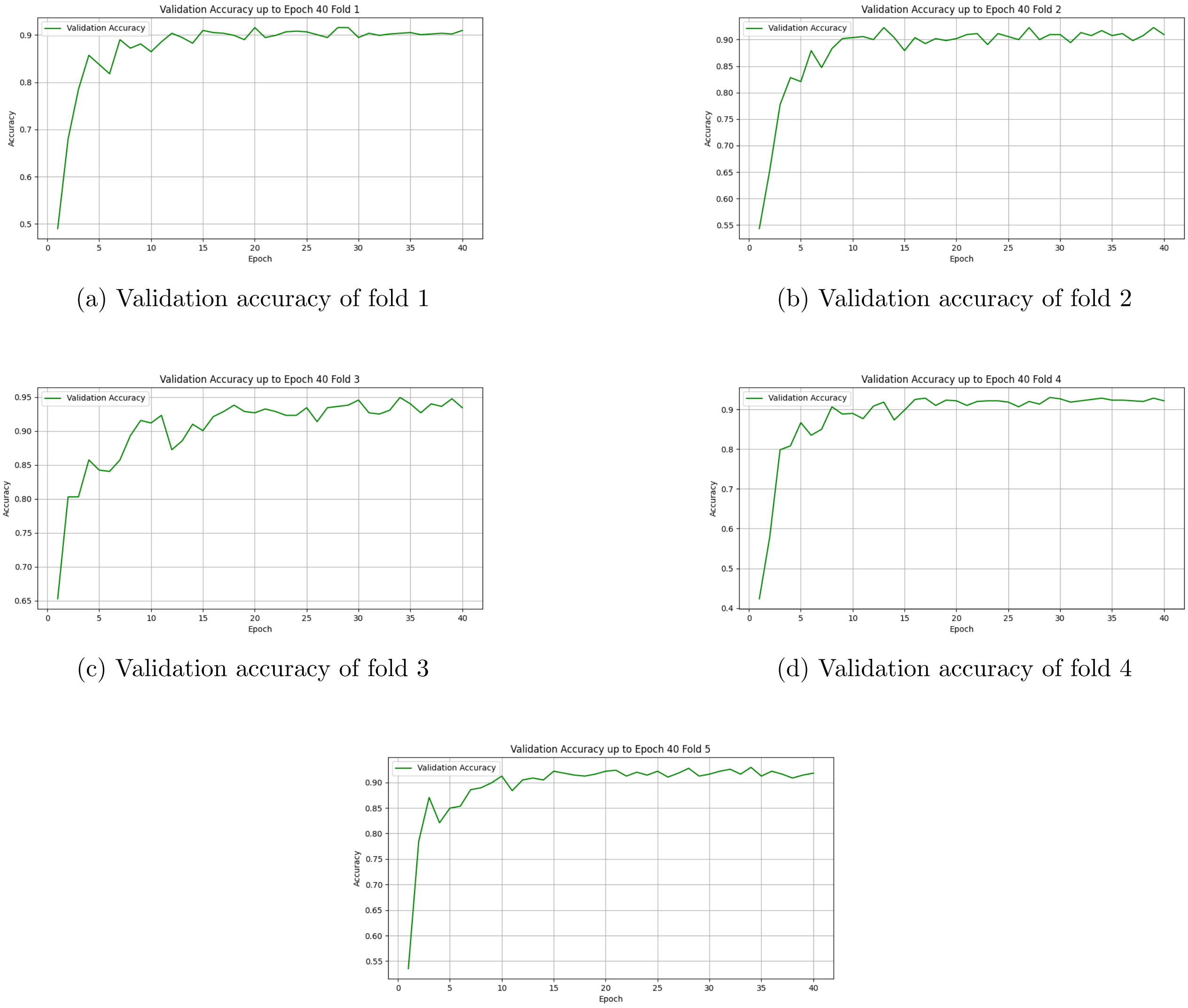
(b) Loss of fold 2

(c) Loss of fold 3

(d) Loss of fold 4

(e) Loss of fold 5

Figure 16: Loss for each fold (ViT and LSTM with Cropped Ankle Images Using Attention Fusion Mechanism of Joint and Image Features)



(a) Validation accuracy of fold 1

(b) Validation accuracy of fold 2

(c) Validation accuracy of fold 3

(d) Validation accuracy of fold 4

(e) Validation accuracy of fold 5

Figure 17: Validation accuracy for each fold (ViT and LSTM with Cropped Ankle Images Using Attention Fusion Mechanism of Joint and Image Features)

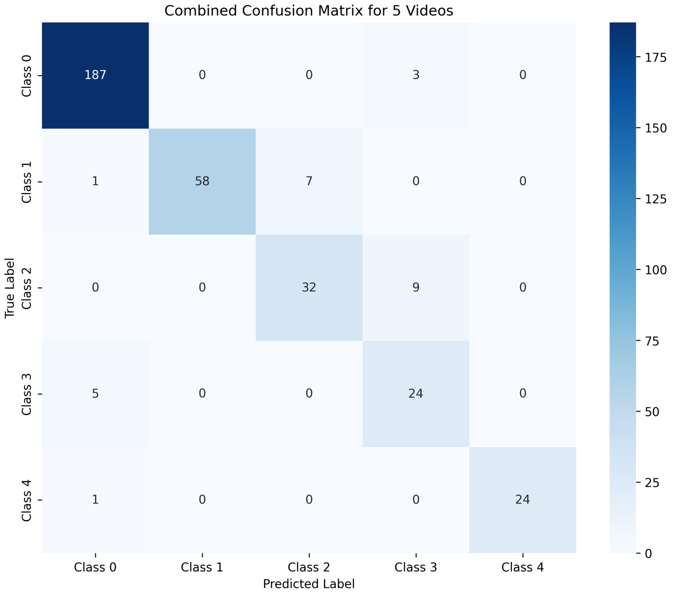
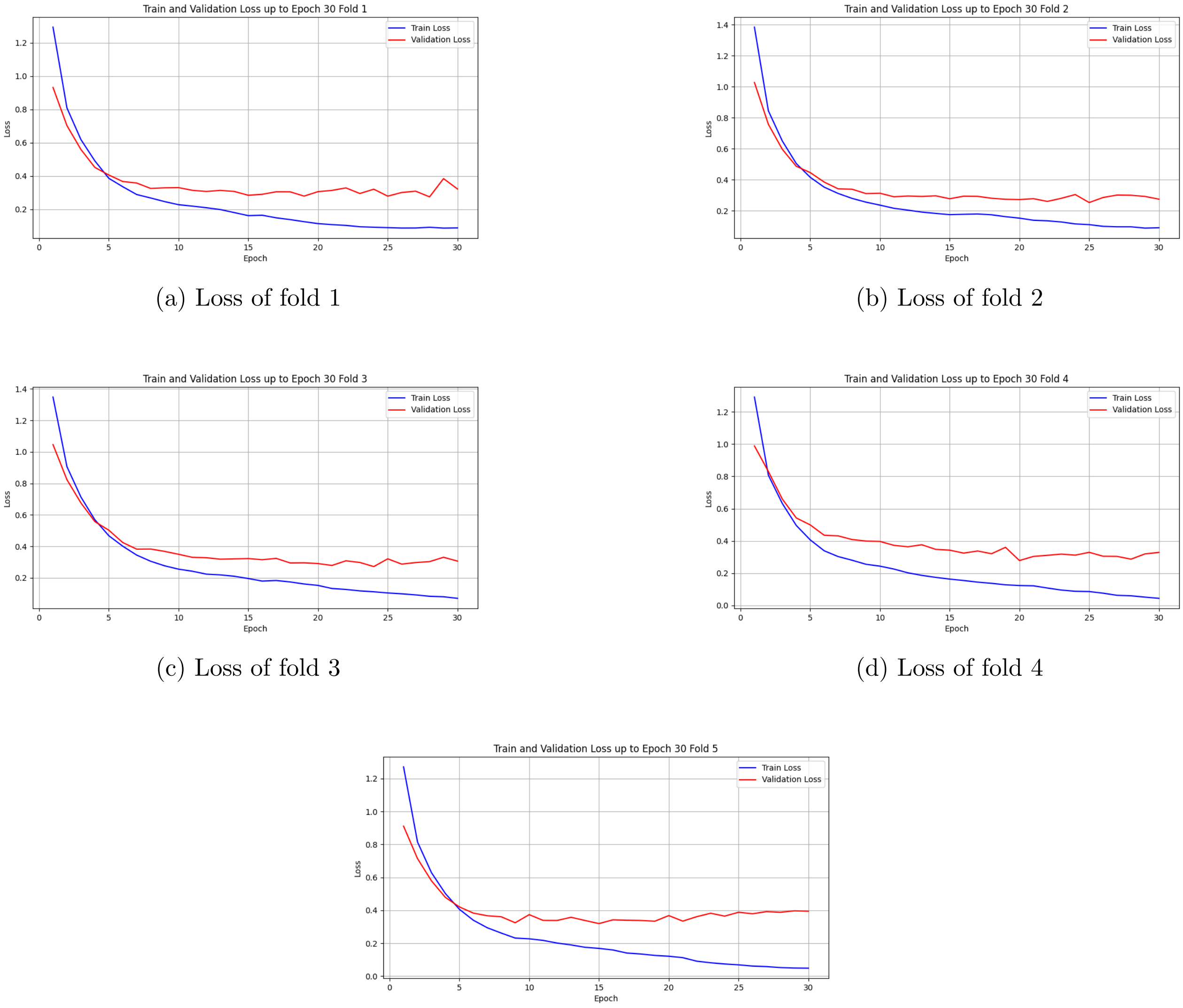


Figure 18: Confusion matrix (ViT and LSTM with Cropped Ankle Images Using Attention Fusion Mechanism of Joint and Image Features)



(a) Loss of fold 1

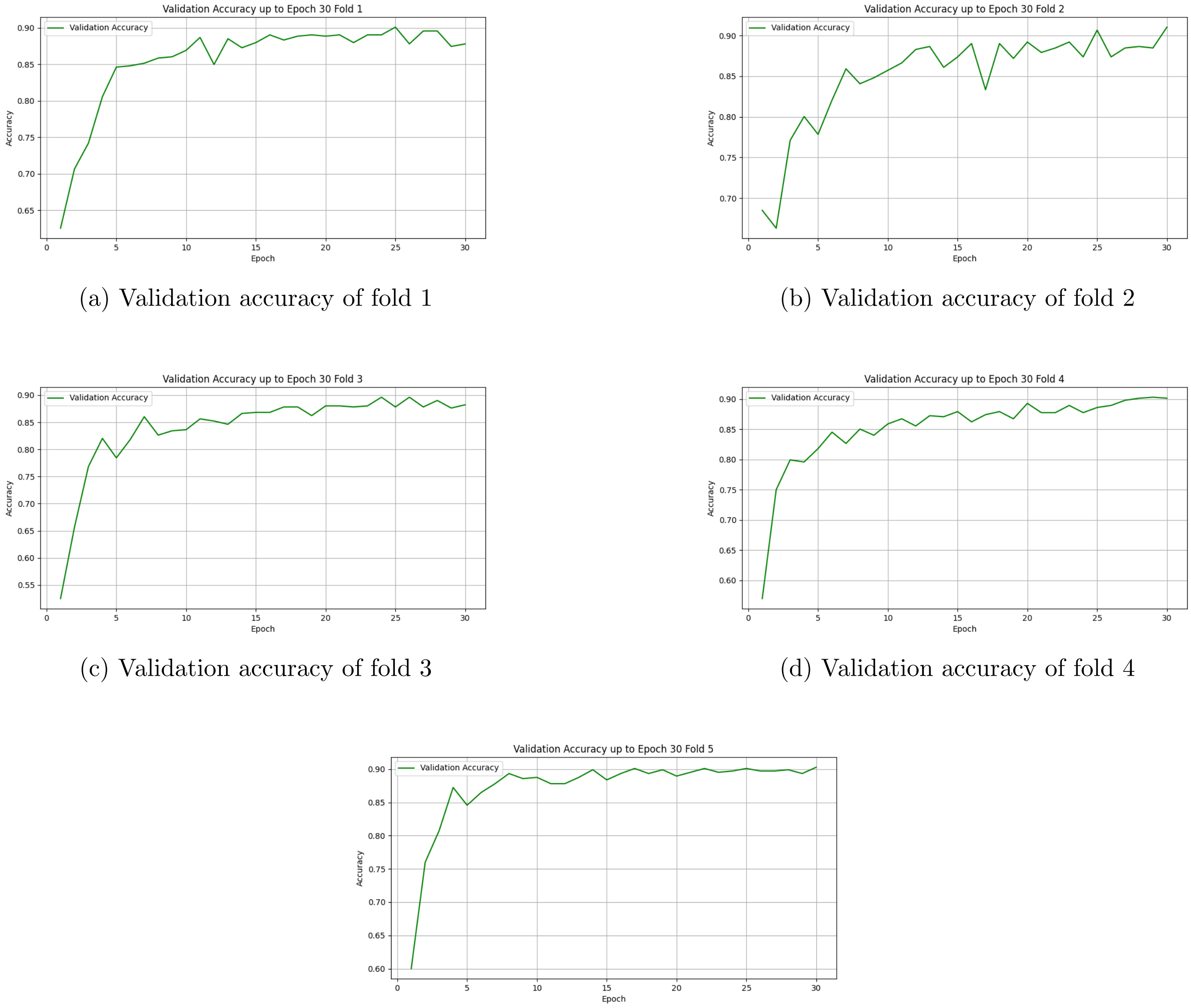
(b) Loss of fold 2

(c) Loss of fold 3

(d) Loss of fold 4

(e) Loss of fold 5

Figure 19: Loss for each fold (ViT and LSTM with Cropped Ankle Images Using Concatenation of Joint and Image Features)



(a) Validation accuracy of fold 1

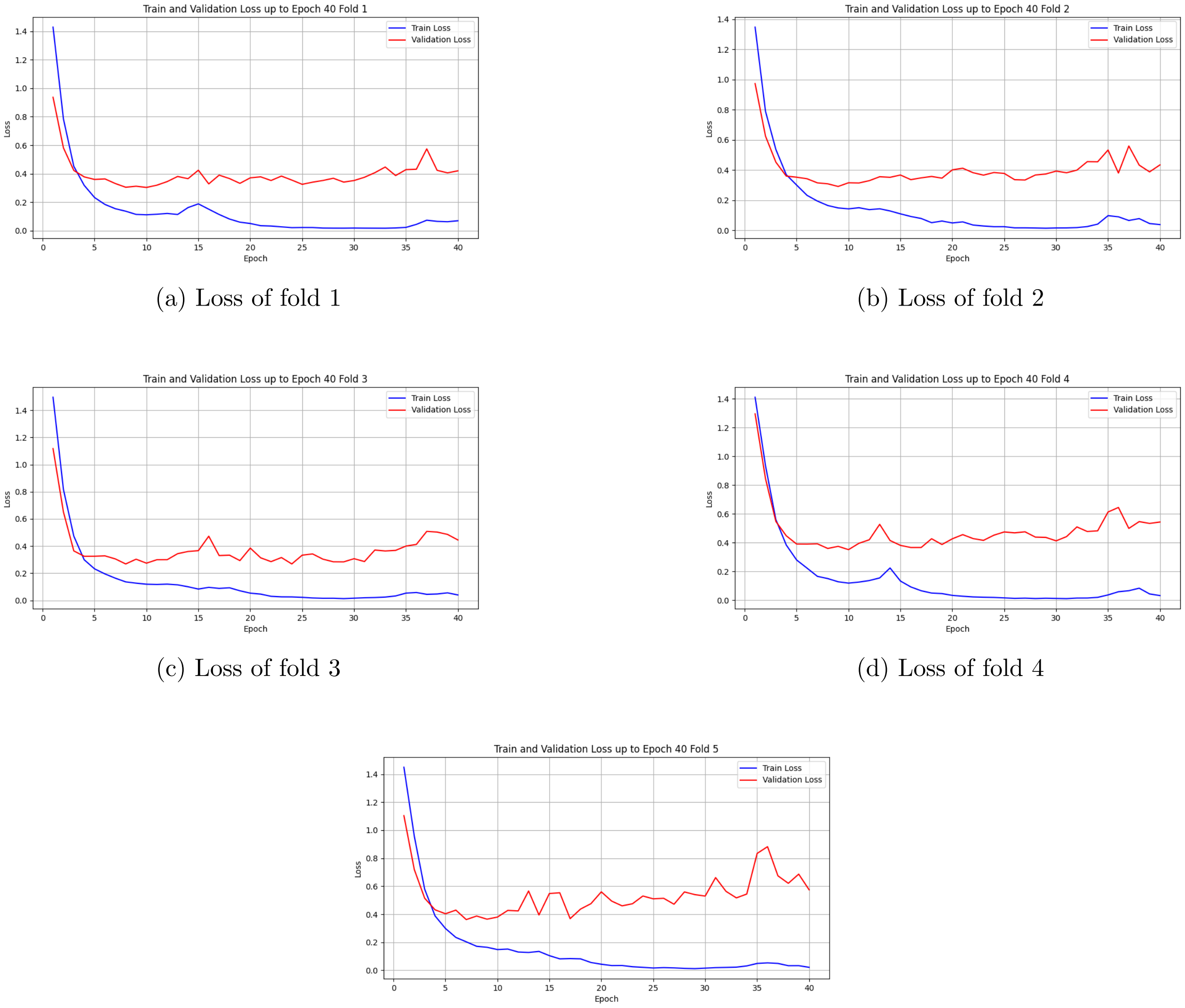
(b) Validation accuracy of fold 2

(c) Validation accuracy of fold 3

(d) Validation accuracy of fold 4

(e) Validation accuracy of fold 5

Figure 20: Validation accuracy for each fold (ViT and LSTM with Cropped Ankle Images Using Concatenation of Joint and Image Features)



(a) Loss of fold 1

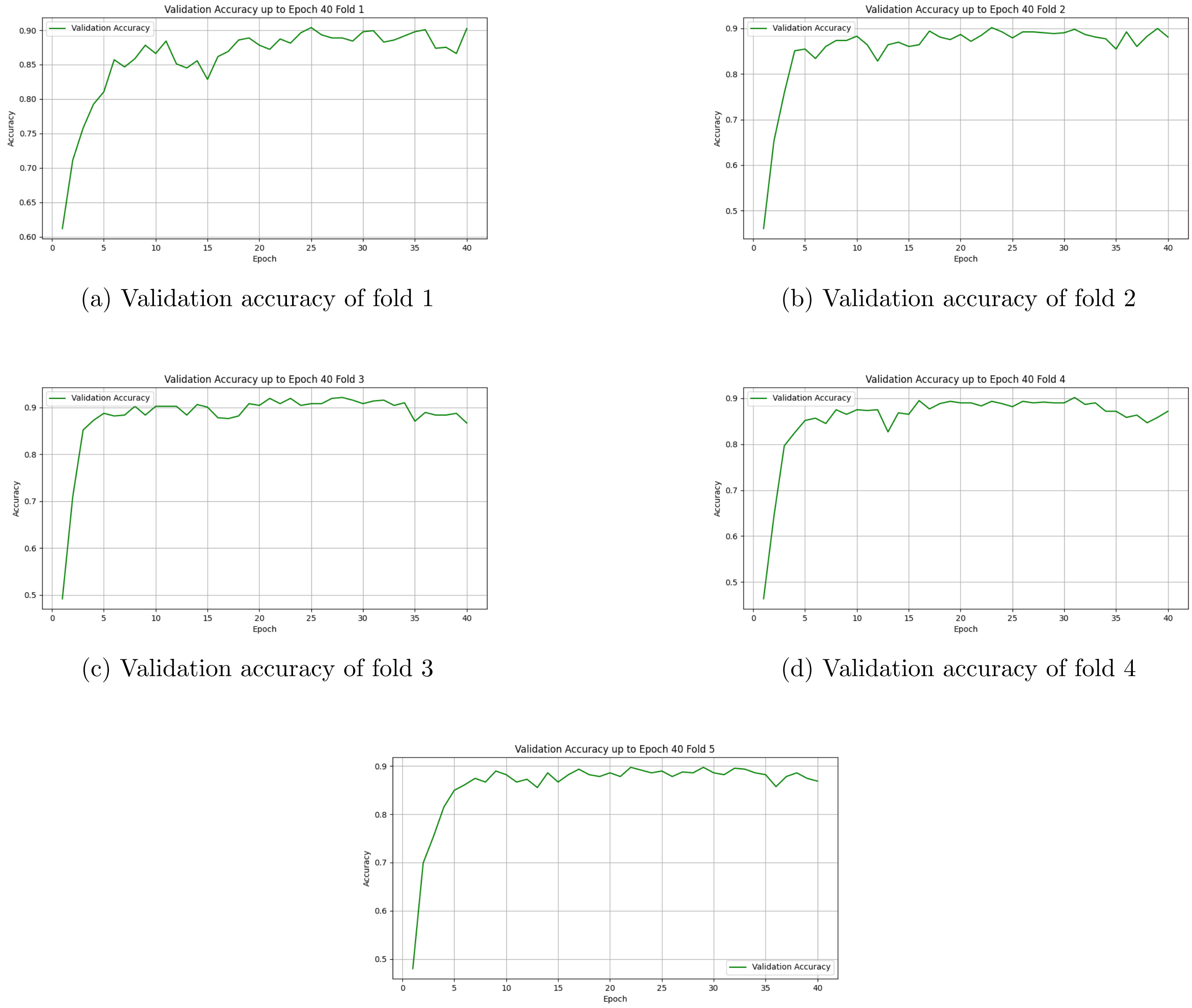
(b) Loss of fold 2

(c) Loss of fold 3

(d) Loss of fold 4

(e) Loss of fold 5

Figure 21: Loss for each fold (ViT and Transformer with Cropped Ankle Images Using Attention Fusion Mechanism of Joint and Image Features)



(a) Validation accuracy of fold 1

(b) Validation accuracy of fold 2

(c) Validation accuracy of fold 3

(d) Validation accuracy of fold 4

(e) Validation accuracy of fold 5

Figure 22: Validation accuracy for each fold (ViT and Transformer with Cropped Ankle Images Using Attention Fusion Mechanism of Joint and Image Features)

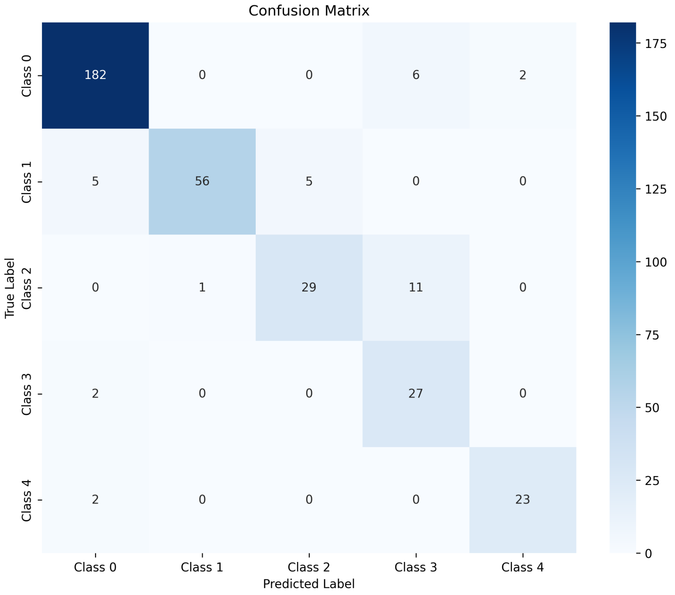


Figure 23: Confusion matrix (ViT and Transformer with Cropped Ankle Images Using Attention Fusion Mechanism of Joint and Image Features)