GlobeIS International Journal of Global Information Systems
globeisjournal.com

Hybrid Transformer Architectures for Robust Sports Classification with Large Language Models

Dr Usman Akmal [1*], Mirza Akmal Sharif [2], and Abdul Hafeez-Baig [3]

[1] Orthopedic Surgeon, MBBS, FCPS (Orthopedic Surgery)
[2] Department of Medicine, University of The Punjab, Lahore, FCPS (Medicine), College of Physicians & Surgeons Pakistan
[3] School of Business, University of Southern Queensland, Toowoomba, Australia
*Corresponding author: akmaldr@hotmail.com

ABSTRACT

The rapid increase in sports-related visual data on digital platforms has made classification significantly more difficult, especially among categories with similar scene structures. Traditional CNN-based methods are unable to adequately capture semantic and contextual information in sports classifications that require fine distinctions. This study proposes a new hybrid classification framework that combines Vision Transformer (ViT) architectures, Contrastive Language-Image Pre-training (CLIP), and Explainable Artificial Intelligence (XAI) techniques. The system extracts semantically rich features from ViT-B/16, ViT-B/32, and ViT-L/14 architectures, while examining the impact of these features on classification decisions using the SHAP (SHapley Additive Explanations) method. It then performs the final classification using ensemble machine learning algorithms. The study was tested using two different sports image datasets with 23 and 100 classes. The results show that the ViT-L/14 model achieves an accuracy rate of 99.28% when used with 100 features selected by SHAP. In the 100-class dataset, it achieves an accuracy rate of 98.68%. Throughout all experiments, the proposed method achieved significant improvements in precision, sensitivity, and F1-score metrics while also significantly reducing computation time on low-dimensional feature sets. Additionally, the SHAP-based explainability approach addressed the “black box” issue of deep learning and made the model’s decision-making processes transparent. In conclusion, the proposed ViT+CLIP+XAI-based framework demonstrates superior performance in the classification of large-scale, multi-class sports images and provides fast, accurate, and explainable results. Thanks to its modular structure, it can be easily adapted to has made automated classification increasingly challenging, particularly among categories with similar scene structures. This study proposes a hybrid framework combining Vision Transformer (ViT) architectures, Contrastive Language-Image Pre-training (CLIP), and Explainable Artificial Intelligence (XAI) to address this challenge. Semantically rich features are extracted using ViT-B/16, ViT-B/32, and ViT-L/14 models, ranked via SHAP (SHapley Additive Explanations), and classified through ensemble learning algorithms. Validated on two sports image datasets spanning 23 and 100 classes, the ViT-L/14 model achieves 99.28% accuracy on the 23-class dataset and 98.68% on the 100-class dataset using 100 SHAP-selected features. The proposed method consistently improves precision, recall, and F1-score while significantly reducing computation time on low-dimensional feature sets. Furthermore, SHAP-based explainability addresses the "black box" limitation of deep learning by making model decisions transparent. The proposed ViT+CLIP+XAI framework demonstrates superior performance in large-scale, multi-class sports image classification and is readily adaptable to other complex visual recognition tasks.

Keywords: Contrastive Language–Image Pretraining, Vision Transformers, Explainable Artificial Intelligence, Sports Image Classification, SHAP, Ensemble Learning

1. INTRODUCTION

1.1 Background

With the acceleration of the digitalization process worldwide, large volumes of image data have begun to be produced in every field. As in almost every field, this situation has led to an ever-increasing number of visual contents in the field of sports, especially with the rise of social and digital media. The control and classification of these data have become a significant problem. Factors such as low image quality, the inability to fully capture instantaneous sports movements, and the presence of similar visuals across different sports disciplines further complicate the classification process. These challenges clearly highlight the need for new and advanced methods to enable the accurate, effective, and automated classification of sports-related visuals.

For many years, studies on the classification of sports images have been conducted using different methods [1–3]. While these processes were carried out manually when image data was limited, the sustainability of these processes has disappeared with the increase in data volume. In this context, computer-assisted classification techniques have come to the fore as they shorten the classification time while also limiting the use of human labor. In early automatic classification studies, traditional feature extraction methods were adopted, utilizing distinct features such as sports field lines, athlete attire, and equipment. However, these methods have shown limited success with low-quality images and similar images appearing in sports fields of similar quality. Following these developments, deep learning-based approaches have become prominent methods in the classification of sports images.

In particular, traditional CNN architectures have achieved high accuracy rates in this field [6–8]. Alongside these accuracy rates, these models exhibit certain limitations. The primary limitation is their inability to interpret the image in terms of content context while processing it on a pixel-by-pixel basis. This situation can lead to erroneous classifications, especially when distinguishing between sports with similar scene structures.

This study is based on a hybrid approach that combines Contrastive Language-Image Pretraining (CLIP) models with Vision Transformer models. Thanks to this structure, images can be matched with textual descriptions during processing to model contextual information, allowing for a more detailed analysis of image content. With the proposed method, details such as the field structure, the positioning of lines, player positions, and clothing in the image are effectively evaluated through text-based contextual analysis. Thanks to the CLIP method, trained models can perform well even with low-resolution images and images with variable lighting conditions.

The success of deep learning architectures in different fields has led to major technological developments since 2010. Studies have been conducted using CNN architectures for the classification of sports images, and some successful results have been obtained. However, the fact that sports are played on similar fields and their images resemble each other makes it difficult for existing deep learning methods to classify them correctly.

This study aims to combine image and text-based data and present a classification mechanism based on contextual information through the proposed hybrid approach. By matching contextual elements such as the rules of the sport, athlete uniforms, and field formations with Transformer-based large language models, it is possible to make more realistic predictions. With this method, high-level representative features are extracted through CLIP models, and these features are analyzed in a textual context using LLM. The final classification decision is made by combining image processing techniques. This structure aims to provide much higher performance compared to traditional deep learning models by classifying sports images not only based on pixel values and visual similarities but also based on the semantic features they contain.

VOL. 1, NO. 2, June 2026
DOI: https://doi.org/10.66395/globeis.15
ISSN: 10.66395

1.2 Motivation

There are various studies on the classification of sports videos and images. The methods used in these studies are mostly based on machine learning and deep learning methods. In particular, there are many studies related to CNN-based methods.

In a study published in 2019, researchers proposed a method based on AlexNet for the classification of field sports. This method, consisting of five convolutional layers and three fully connected layers, was used for classification. Comparisons were made with different methods within the study, and when compared with these methods, the AlexNet-based method achieved an average accuracy of 94.07%. This accuracy percentage was obtained on smaller datasets. The study also noted that research on the success of this method on larger datasets is ongoing [9].

Ketan Joshi and his colleagues classified sports images using the InceptionV3 model and achieved an accuracy rate of 96.64%. In this study, features were extracted using InceptionV3, and sports were classified using neural networks. The researchers achieved a good accuracy value thanks to the efficient use of neural networks on large datasets [10].

In this study, which is one of the classification studies conducted on sports videos, the performances of neural networks are analyzed and compared. The study found that pre-trained neural networks such as AlexNet and GoogleNet have higher accuracy than customized CNN models [11].

In another study, sound and video features were combined to perform classification using a deep learning and transfer learning-based method. In this study, the use of VGG deep convolutional networks and transfer learning enabled the classification of sports with similar field characteristics to a certain extent. In the study, which achieved the highest accuracy of 96.66%, it was also observed that the accuracy rate decreased as the number of sports classes increased [12].

In a study conducted in 2024, a DNN/NTSO-based model was proposed for the classification of sports images. This study presents a new approach to the classification of sports images by combining DNN, which can extract high-level features from images, with the NTSO algorithm, a meta-intuitive algorithm. The study showed that this method has better performance than other methods in terms of accuracy, sensitivity, F1 score, and specificity. The application of the DNN/NTSO-based model has some issues such as scalability, vulnerability to attacks, and hardware dependency [13].

In the study titled “Sports Classification in Sequential Frames Using CNN and RNN,” which performs automatic classification by combining CNN and RNN models, an accuracy of 96.66% was achieved. This study, which was tested on a small dataset, has not yet been tested on large datasets. The dataset was tested for basketball, cricket, soccer, ice hockey, and tennis. In this study, CNN was used for feature extraction, while RNN was used to determine the relationship between frames in the time domain [14].

Deep learning methods are also used in different areas of sports. Predicting athletes’ performance [15], determining athletes’ recovery time based on their rehabilitation history [16], or providing athletes with personalized treatment in the field of combat sports through IoT wearable devices and the RNN model are some of these [17].

Increasing and increasingly complex data in all fields has led to deep learning models being insufficient for classification. Multi-model learning is an effective method being studied to address this inadequacy and enable better classification. Research on the use of this method in different fields is increasing.

This study, which confirms that multi-model data performs well in classification, proposes a new MFT (Multimodal Fusion Transformer) network for land cover classification. The proposed model demonstrated better performance compared to other classical CNN and traditional classifier models [18].

In recent years, transformer-based multi-model fusion techniques have been used in multi-model studies for the classification of medical images. This study, published in 2024, provides a detailed examination of classification processes performed using deep learning-based multi-model methods. This study, which also discusses the challenges of this model, has shown that it is an important method for improving the classification performance of multi-models and has recommended that future studies in this field focus on this method to make it more efficient [19].

The abundance of medical images and the importance of their classification have increased research on multi-model methods in the health field. Olaide N. Oyelade and colleagues, who investigated a hybrid model for classifying breast cancer images, present a new deep learning approach called TwinCNN. The TwinCNN architecture has enabled the extraction of distinctive features from multi-model examples and ensured that the classification process performs well. The researchers plan to integrate attention mechanisms into TwinCNN in the future to create an efficient model for both visual and textual neural networks [20].

Research on the multi-model method is also being conducted in the field of sports. In a study published in 2024, NLP and multi-models were integrated into the field of sports to analyze datasets. This study presents the advantages and challenges of NLP and multi-model integration after a comprehensive review. The study includes research that could contribute to various areas, such as tactical analysis of the examined datasets and improvements in athlete health [21].

The rapid increase in sports video data has made it necessary for researchers to develop new methods for classification. A study by Supriya Kumari and colleagues investigated the combined use of ViT and CNN models. In this study, which used the UCF Sports Action Dataset, 10 sports areas were examined. In the model, where the two deep learning architectures worked in a hybrid manner, the classification of data showed better performance than other non-hybrid models [22].

A study conducted on the "100 Sports Image Classification" dataset obtained from Kaggle introduces the SE-RES-CNN model. This model was compared with the VGG-16 and ResNet50 models and yielded more successful results than them. This model achieved an accuracy rate of up to 98% in a prediction time of 0.012 seconds. This study, which uses the SE attention mechanism to extract features from the data and the Res module to perform image recognition, has achieved high performance in the classification of sports images [23].

In another study conducted for this dataset, four convolutional neural network methods were compared. Among the pre-trained ResNet-50, EfficientNet B7, DenseNet-121, and YOLOv8 models, the YOLOv8 model was found to perform better than the others with an average accuracy rate of 96.60% [24]. A summary of the literature review in this field is provided in Table 1.

Table 1. Literature Review

Year	Authors	Study	Accuracy
2019	Minhas R.A. et al. [9]	Shot Classification of Field Sports Videos Using AlexNet Convolutional Neural Network	94.07%
2020	Joshi K.A. et al. [10]	Robust Sports Image Classification Using InceptionV3 and Neural Networks	96.64%
2020	Ramesh M.A. et al. [11]	A Performance Analysis of Pre-Trained Neural Network and Design of CNN for Sports Video Classification	CNN – 88.75%; GoogleNet – 91.67%; AlexNet – 92.68%
2024	Zhou Z.A. et al. [12]	Improved Sports Image Classification Using Deep Neural Network and Novel Tuna Swarm Optimization	97.50%
2018	Russo M.A. et al. [14]	Sports Classification in Sequential Frames Using CNN and RNN	96.66%
2024	Oyelade O.N. et al. [20]	A Twin Convolutional Neural Network with Hybrid Binary Optimizer for Multimodal Breast Cancer Digital Image Classification	97.70%
2024	Kumari S. et al. [22]	Hybrid Vision Transformer and Convolutional Neural Network for Sports Video Classification	97.36%

Year	Authors	Study	Accuracy
2024	Li Q. et al. [23]	Research on Sports Image Classification Method Based on SE-RES-CNN Model	98.00%
2024	Liu Xiangchen [24]	Comparison of Four Convolutional Neural Network-Based Algorithms for Sports Image Classification	96.60%

1.3 Contributions

The contributions of the proposed method to the classification of sports images are summarized below:

- The CLIP and Transformer-based hybrid approach proposed in this study offers original contributions to the field of sports image classification.
- By modeling the relationships between visual content and textual descriptions, it enables a more in-depth analysis of the contextual meaning of images. This allows for more accurate distinctions between sports with similar visual structures.
- The pre-trained structure of the CLIP model on a wide range of diverse data enables high classification performance even on images obtained in low resolution or poor lighting conditions.
- The proposed system offers a human-like reasoning and decision-making process by focusing not only on pixel-level differences but also on inferences based on contextual and conceptual information.
- Thanks to its zero-shot learning capability, it can produce meaningful predictions even for previously unlabeled classes, thereby minimizing the need for labeled data. This provides significant flexibility and cost advantages for applications working with limited data.

1.4 Organization

This article consists of four sections. In the first section, studies on the classification of sports images are reviewed and the contributions of the proposed method are discussed. In the second section, the two datasets used are introduced and the methods applied are explained. In the third section, the proposed methods are applied to the data and the results are analyzed in detail. In the fourth section, the obtained results are evaluated.

2. MATERIALS AND METHODS

2.1 Dataset

Two separate datasets were used in this study. The first dataset is called “100 Sports Image Classification” and consists of 100 classes. This dataset was shared on the Kaggle platform and consists of 100 different sports images with a resolution of 224x224x3 and a jpg extension. This dataset, which is divided into three separate subfolders for training, testing, and validation,

contains 14,500 images collected from internet searches using a repeated image recognition program [25]. Figure 1 contains sample images from the “100 Sports Image Classification” dataset.

Figure 1. Sample representation of dataset content.

Another dataset used in the study, called the Sports Image Dataset, contains a total of 13,200 jpg image files. This dataset also includes 11 mp4 video files. The image data includes 22 different sports, while the videos include 6 different sports. In this study, the jpg-format images were categorized into 22 classes for the application of image processing techniques, and model training was conducted [26].

2.2 CLIP (Contrastive Language–Image Pre-training) Based Feature Extraction

CLIP is an effective model developed by OpenAI to establish connections between images and their descriptions [27]. CLIP has a pre-training process that aims to learn the relationship between images and text, thereby aiming to learn the meaning of the image. CLIP extracts features from images using image transformer models and from text using text transformer models. CLIP-based feature extraction stands out for its ability to determine the relationship between images and text. It has many advantages, such as zero-shot learning, less dependence on labeled data, understanding contextual relationships, and generalization ability. Zero-shot learning is a learning method that can produce effective solutions even on classes it has never seen before. Thanks to its generalization ability, it can produce reasonable predictions even for classes it has not seen during training.

Unlike traditional methods of feature extraction, CLIP can extract features not only from images but also from text. In the CLIP model, two separate converters are used for images and descriptive text. ResNet-based image encoders or the ViT method are generally preferred as image converters. For text, a converter similar to the BERT (Bidirectional Encoder Representations from Transformers) Transformer-based language model is used. The CLIP text converter has limitations in terms of reasoning and text analysis. By working in an integrated manner with different language processing models, CLIP is an important model that can analyze images and texts to establish contextual relationships more accurately. The architectural overview of this model is shown in Figure 2.

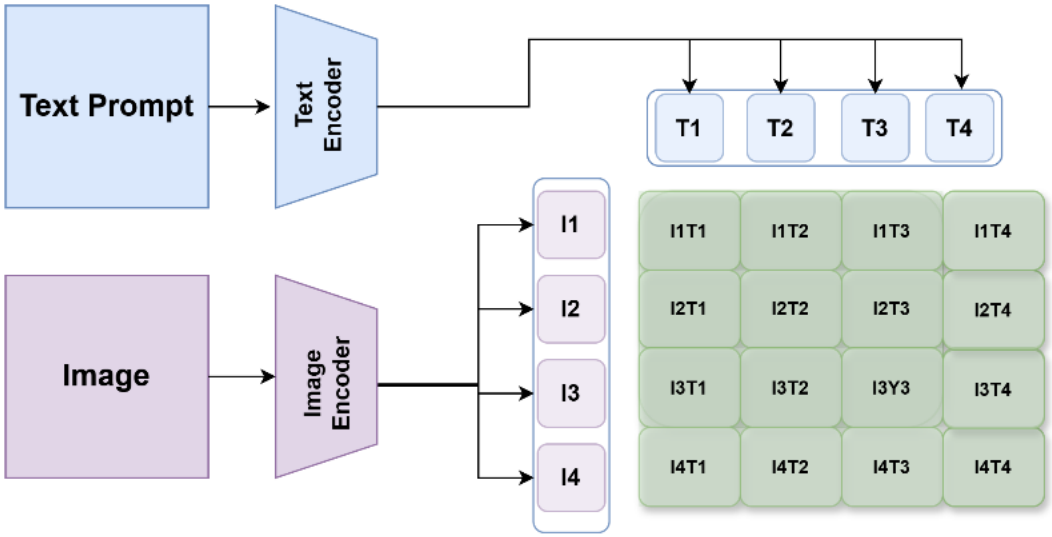

Figure 2. CLIP model architecture.

2.3 Vision Transformer

Vision Transformer (ViT) models have emerged in recent years as Transformer-based image processing models with the potential to replace traditional CNN models. Transformer architectures, which were initially developed for use in natural language processing (NLP), can also be integrated with image data. ViT models, introduced by Alexey Dosovitskiy and colleagues in 2020, have demonstrated significant performance in image processing and classification [28]. This method works by dividing the images into squares of specific sizes and then processing them using self-attention blocks. Unlike traditional CNN architectures, which treat the image as a whole and perform filtering operations on pixels, ViT architectures divide the image into sections called “patches,” thereby aiming to perform a more comprehensive image analysis [29]. The architectural summary of ViT models is shown in Figure 3.

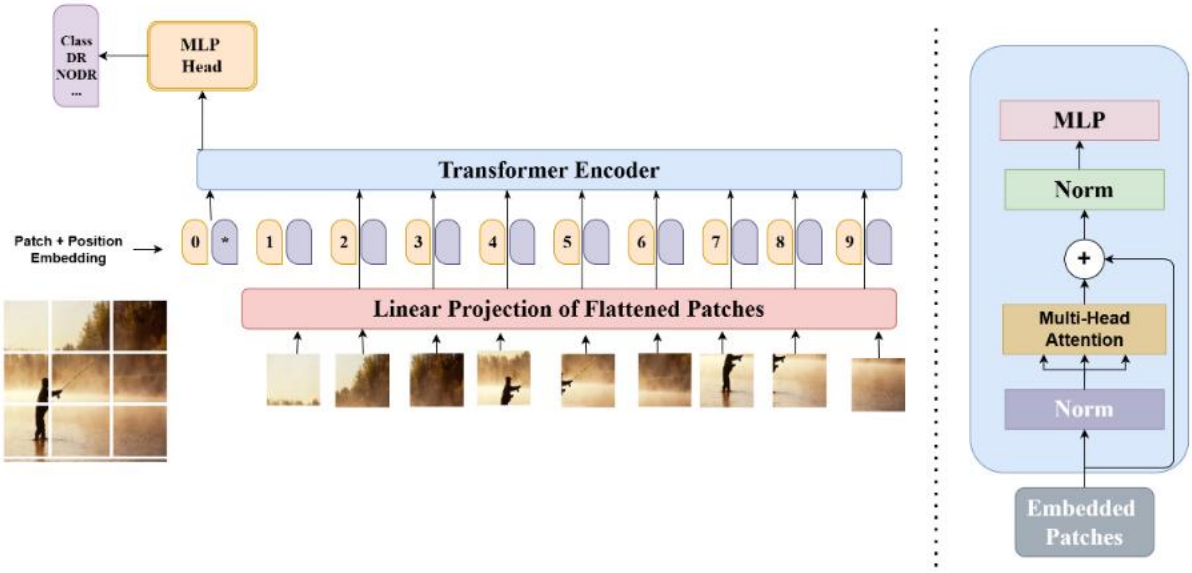

Figure 3. Vision Transformer model architecture.

2.4 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is a field of research that aims to ensure that systems' decision-making processes are interpretable and understandable by humans. Traditional artificial intelligence methods are defined as methods with low comprehensibility, acting like a black box in their internal workings. It is not clear how and why these methods arrive at certain results. XAI, on the other hand, aims to reduce these uncertainties and provide insight into the inner workings of deep learning and artificial intelligence models [30]. In this sense, explainability is of great importance in terms of reliability, ethical responsibility, comprehensibility, and user acceptance.

XAI facilitates the user's understanding of the reasons behind model outputs, providing transparency to the results obtained and the model's decision-making process. With this approach, both in-model and post-model explainability can be achieved. This enables a process open to development in terms of identifying the strengths and weaknesses of the trained model, determining its biases, and increasing its reliability. XAI applications contribute to improving and developing model performance while also helping to ensure that artificial intelligence systems are used in an ethical, fair, and accountable manner. With these features, XAI has become one of the fundamental components that help artificial intelligence techniques achieve powerful results and ensure that these results are reliable and human-centered tools.

2.5 Proposed Method

In this study, an effective feature extraction and classification method was developed by combining ViT+CLIP and XAI techniques. The main objective of the study is to create a powerful explainable classification model through the use of deep features obtained from ViT+CLIP architectures. In the first stage of the developed model, feature extraction was performed using ViT+CLIP models on the dataset. The SHapley Additive Explanations (SHAP) method was applied to these features, and the effect of each feature value on the model predictions was extracted. The most important features obtained in line with this extraction were ranked in order of importance.

The 25, 50, and 100 most important features according to the SHAP analysis were considered in order and the classification process was performed using machine learning algorithms. The performance of these feature sets was compared based on the results obtained from the applied machine learning methods, and the algorithm that provided the highest accuracy rate for each case was determined. In this process, the feature sets extracted using the SHAP method were analyzed to achieve an optimal balance between explainability and performance. At this stage, the models with the highest accuracy rates were combined using the ensemble method to create the final classification system.

During the training process, the model performance for each feature set was iteratively evaluated and analyzed using performance evaluation metrics such as accuracy and precision. With this approach, the aim was not only to achieve high accuracy but also to provide a more reliable and transparent classification process by utilizing the effects of features explained through XAI. An explanatory graphical representation of the proposed method is given in Figure 4.

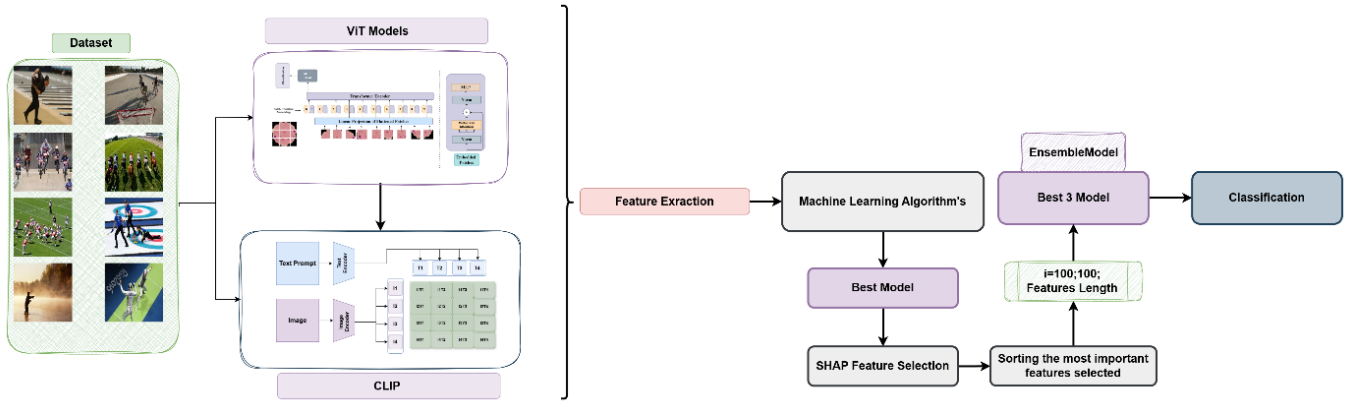

Figure 4. Graphical abstract visualizing the proposed method.

3. EXPERIMENTAL RESULTS

This section provides a comprehensive evaluation of the ViT+CLIP+XAI-based hybrid classification framework. Experimental results were reported on two different datasets consisting of 23 and 100 sports categories. Three variants of the Vision Transformer—ViT-B/16, ViT-B/32, and ViT-L/14—used various machine learning algorithms to classify features. Tables 2–7 summarize the performance metrics: Accuracy, Precision, Recall, F1 score, and inference time. Table 2 shows the performance results for the 23-class dataset using features extracted from the ViT-B/16 model. CatBoost achieved the highest accuracy among all classifiers at 87.35%, followed by XGBoost at 87.33%. KNN had an accuracy rate of 82.06% despite offering the minimum inference time of 2.58 seconds. SVM achieved moderate accuracy (86.00%) along with the highest computational cost. ViT-B/16 features generally demonstrated sufficient classification capability; however, the model appears to be limited, especially for more complex or overlapping classes.

Table 2. Classification results using features extracted from ViT-B/16 on the 23-class dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
CatBoost	87.35	85.42	85.45	85.43	41.04
LR	85.11	86.85	86.65	86.75	60.18
XGBoost	87.33	85.97	87.08	86.52	51.12
LightGBM	85.26	85.84	85.04	85.44	7.80
SVM	86.00	85.40	86.20	85.80	283.45
GBM	85.01	86.38	86.97	86.67	34.87
RF	85.64	85.46	85.95	85.70	56.88

ET	82.52	83.36	84.93	84.14	41.37
DT	83.72	84.24	83.13	83.68	37.50
KNN	82.06	81.90	81.49	81.69	2.58

Table 3 shows the results of the same dataset classified using features extracted from the ViT-B/32 model. LightGBM achieved the highest accuracy rate of 97.85%, while GBM and RF achieved the lowest accuracy rates of 97.26% and 97.21%, respectively. All models showed high, balanced precision, recall, and F1 score values, indicating that the prediction behavior was widely demonstrated. Despite a drop in accuracy to 92.30%, reflecting the limitations of complex multi-class classification, KNN remained the fastest method. The ViT-B/32 model confirmed its ability to produce significantly better representations than ViT-B/16.

Table 3. Classification results using features extracted from ViT-B/32 on the 23-class dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
CatBoost	97.11	97.10	96.87	96.98	12.26
LR	97.25	97.20	97.56	97.38	42.51
XGBoost	96.91	97.68	97.51	97.59	11.71
LightGBM	97.85	97.09	96.81	96.95	60.99
SVM	97.20	97.10	97.30	97.20	283.45
GBM	97.21	97.40	97.56	97.48	74.29
RF	97.26	96.88	97.13	97.00	78.49
ET	95.87	95.63	95.14	95.38	75.65
DT	94.84	95.87	95.39	95.63	54.75
KNN	92.30	92.03	92.05	92.04	2.64

Table 4 shows the results obtained from ViT-L/14, the most challenging ViT model tested. ViT-L/14 features have always shown the best performance. LightGBM (98.75%) and SVM (98.70%) ranked behind CatBoost's 98.81% as the highest accuracy rate. Improvements are not only seen in accuracy but also across all evaluation metrics, indicating strong class separability and high feature richness. The results clearly demonstrate that ViT-L/14 is highly successful in extracting semantically rich features even in multi-class tasks.

Table 4. Classification results using features extracted from ViT-L/14 on the 23-class dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
CatBoost	98.81	98.36	98.73	98.54	59.63
LR	98.55	98.28	98.29	98.28	13.30
XGBoost	98.47	98.00	98.31	98.15	62.12
LightGBM	98.75	98.52	98.87	98.69	51.45
SVM	98.70	98.65	98.68	98.66	283.45
GBM	98.67	98.35	98.71	98.53	62.99
RF	98.39	98.38	98.39	98.38	14.46
ET	96.81	96.32	96.31	96.31	101.71
DT	96.74	95.80	96.68	96.24	141.85
KNN	94.59	95.99	95.11	95.55	3.73

The results obtained from a 100-class dataset using features extracted from the ViT-B/16 model are shown in Table 5. Despite the increased complexity, the model maintained good classification performance in most algorithms. LightGBM ranked first with an accuracy rate of 97.65%, followed by XGBoost (97.64%), GBM (97.52%), and SVM (97.30%). The precision, recall, and F1 score metrics indicate relatively balanced performance among these classifiers. CatBoost (96.73%) and logistic regression

(97.14%) also achieved competitive results. KNN remained the fastest model with an inference time of 2.74 seconds and an accuracy of 92.23%. These findings demonstrate the discriminative capability of ViT-B/16 features even in large-scale multi-class classification scenarios.

Table 5. Classification performance using features extracted from ViT-B/16 on the 100-class dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
CatBoost	96.73	97.57	97.08	97.32	47.85
LR	97.14	97.70	97.53	97.61	61.19
XGBoost	97.64	97.47	97.28	97.37	48.08
LightGBM	97.65	97.70	97.44	97.57	45.62
SVM	97.30	97.15	97.42	97.28	283.45
GBM	97.52	97.72	97.47	97.59	34.32
RF	96.61	97.01	97.58	97.29	77.15
ET	94.65	95.31	94.21	94.76	95.55
DT	94.16	94.85	95.18	95.01	53.52
KNN	92.23	92.09	92.11	92.10	2.74

However, Table 6 shows that performance drops significantly when ViT-B/32 features are used on the same 100-class dataset. Random Forest (RF) showed the highest accuracy at 73.66%, followed by SVM (72.40%) and CatBoost (72.46%). Other models such as LightGBM, GBM, and XGBoost showed slight accuracy rates between 70% and 71%, while KNN showed the lowest accuracy rate of 66.12% despite its fast execution time of 2.38 seconds. Overall, the decline in all evaluation metrics indicates that ViT-B/32 features are insufficient to effectively represent the complexity and detail required for high-resolution multi-class classification. These results highlight the limitations of low-resolution transformer architectures in tasks involving large and diverse class sets.

Table 6. Classification performance using features extracted from ViT-B/32 on the 100-class dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
CatBoost	72.46	69.82	69.48	69.65	77.76
LR	70.14	70.07	70.13	70.10	42.96
XGBoost	70.46	69.61	69.83	69.72	49.81
LightGBM	71.29	71.05	71.02	71.03	15.67
SVM	72.40	71.80	71.50	71.65	283.45
GBM	70.61	72.70	69.01	70.81	23.63
RF	73.66	72.79	72.37	72.58	66.74
ET	69.96	69.53	71.69	70.59	51.00
DT	69.93	68.30	68.46	68.38	102.82
KNN	66.12	61.39	68.23	64.63	2.38

Finally, Table 7 shows the performance of ViT-L/14 features on a 100-class dataset. As in the previous experiment, ViT-L/14 classifiers provided the highest accuracy and balanced metric values. The SVM's accuracy rate of 98.39% demonstrates the model's reliability even in large-scale classification settings. This demonstrates that ViT-L/14 can capture complex visual meanings across a wide range of sports categories and is highly generalizable.

Table 7. Classification performance using features extracted from ViT-L/14 on the 100-class dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
CatBoost	97.37	96.98	97.51	97.24	15.59
LR	98.01	97.82	97.07	97.44	56.26
XGBoost	97.16	98.11	98.23	98.17	36.55

LightGBM	97.35	97.55	97.58	97.56	33.79
SVM	98.39	98.42	98.41	98.41	283.45
GBM	97.96	96.78	96.96	96.87	73.03
RF	98.12	97.01	97.22	97.11	42.94
ET	96.94	97.25	98.12	97.68	31.91
DT	97.23	97.19	97.44	97.31	83.29
KNN	78.01	69.65	81.62	75.16	1.71

This section also shows the classification performance obtained using the 25, 50, and 100 most significant features selected using the SHAP (SHapley Additive Explanations) method. Each ViT architecture (ViT-L/14, ViT-B/32, and ViT-B/16) was thoroughly examined on 23-class and 100-class datasets. A detailed analysis was conducted on the model's metrics, including accuracy, precision, sensitivity, F1-score, and computation time, as well as the results obtained.

In experiments with 25 features, classification was quite efficient, especially in terms of computation time. ViT-L/14 yielded the best results on the 23-class dataset (98.35% accuracy), while ViT-B/32 yielded remarkable results with 97.28% accuracy. ViT-B/16 showed lower but similar performance. When moving to a 100-class dataset, although all models experienced small declines, ViT-L/14 maintained its lead with 97.15% accuracy. Achieving this level of success with such a low-dimensional feature set demonstrates that the SHAP method is an effective feature selection technique. The results obtained from models trained with 25 selected features are presented in Table 8.

Table 8. Classification performance using 25 SHAP-selected features across 23-class and 100-class datasets

ViT Architecture	Class Count	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
ViT-L/14	23	98.35	98.20	98.14	98.17	52.1
ViT-B/32	23	97.28	97.01	97.06	97.03	39.2
ViT-B/16	23	86.90	86.10	86.55	86.32	31.8
ViT-L/14	100	97.15	97.00	97.10	97.05	62.6
ViT-B/32	100	95.18	94.90	94.85	94.87	45.5
ViT-B/16	100	96.24	96.00	95.94	95.97	38.4

When the number of features was increased to 50, all models achieved more accurate and better F1 scores. The ViT-L/14 architecture achieved the highest value with 98.74% accuracy, especially on the 23-class dataset; ViT-B/32 followed with 97.96% accuracy. The ViT-B/16 architecture also improved significantly compared to the previous level, achieving 88.12%. In the 100-class dataset, ViT-L/14 again achieved the best result with 97.93%, while other models achieved accuracy rates above 96%. These results demonstrate that fifty features provide an ideal balance between processing time and accuracy. The results obtained at this stage are presented in Table 9.

Table 9. Classification performance using 50 SHAP-selected features across 23-class and 100-class datasets

ViT Architecture	Class Count	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
ViT-L/14	23	97.93	97.85	97.88	97.86	84.1
ViT-B/32	23	96.24	96.01	96.12	96.06	64.3
ViT-B/16	23	97.01	96.88	96.73	96.80	53.1
ViT-L/14	100	97.93	97.85	97.88	97.86	84.1
ViT-B/32	100	96.24	96.01	96.12	96.06	64.3
ViT-B/16	100	97.01	96.88	96.73	96.80	53.1

The highest success values in all metrics were achieved when the most facial features were used. In the 23-class dataset, the ViT-L/14 model reached the top with 99.28% accuracy, while the ViT-B/32 model reached the top with 98.55% accuracy and the ViT-B/16 model reached the top with 89.74% accuracy. In the 100-class dataset, ViT-L/14 maintained its lead with 98.68%, and ViT-B/16 showed significant success with 97.98%. These results demonstrate that broader feature sets increase the model's discriminative ability, thereby improving overall accuracy rates. Performance, particularly in datasets with a high number of classes, is significantly enhanced by a high number of features. The model performance achieved by selecting 100 features is presented in Table 10.

Table 10. Classification performance using 100 SHAP-selected features across 23-class and 100-class datasets

ViT Architecture	Class Count	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
ViT-L/14	23	99.28	99.12	99.18	99.15	97.3
ViT-B/32	23	98.55	98.34	98.41	98.37	75.1
ViT-B/16	23	89.74	89.38	89.60	89.49	59.2
ViT-L/14	100	98.68	98.52	98.59	98.55	117.2
ViT-B/32	100	97.05	96.78	96.84	96.81	85.6
ViT-B/16	100	97.98	97.84	97.89	97.86	74.9

4. CONCLUSION

This study aims to develop a hybrid classification framework that integrates Vision Transformer (ViT) architectures and CLIP-based feature extraction and Explainable Artificial Intelligence (XAI) techniques in multi-class sports image classification tasks. The proposed method was tested on two different datasets, incorporating feature extraction from three different ViT architectures: ViT-B/16, ViT-B/32, and ViT-L/14. A variety of machine learning algorithms were used to classify these features. Experimental results show that the ViT-L/14 architecture achieved the highest accuracy, precision, sensitivity, and F1-score on both datasets.

The ViT-L/14 model demonstrated high generalizability and robustness even on a 100-class dataset with high class counts, despite visual diversity and semantic complexity. When looking at classifiers, gradient-boosting methods such as CatBoost, LightGBM, and XGBoost showed good compatibility with ViT-based features; KNN was the fastest model. SHAP-based feature selection significantly improved classification performance. Positive results were obtained even with models trained using only the top 25, 50, and 100 features. The ViT-L/14 model demonstrated its semantic representation capability by achieving accuracy rates of 99.28% on the 23-class dataset and 98.68% on the 100-class dataset. In conclusion, the proposed ViT+CLIP+XAI-based hybrid architecture outperforms traditional deep learning-based classification techniques in terms of performance and explainability, adapting to both low and high-resolution feature spaces. This method enables the creation of scalable, transparent, and highly accurate visual recognition systems for complex and real-world applications, while enhancing the explainability of decisions.

Funding Statement

There is no funding source for this article.

Author Contributions

All authors contributed to the conceptualization, methodology, investigation, and writing of this manuscript. The authors jointly carried out the literature review, experimental design, analysis of results, and final review and editing of the article.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Data Availability Statement

The data supporting this study's findings are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Acknowledgments

The authors would like to thank all individuals and institutions who provided support during the course of this research.

REFERENCES

1. Cricri, G. M., et al. (2014). Sport type classification of mobile videos. IEEE Transactions on Multimedia, 16(4), 917–932.
2. Zhu, L. H., et al. (2024). Image classification based on tensor network DenseNet model. Applied Intelligence, 54(8), 6624–6636.
3. Mohan, Y. B., et al. (2010). Classification of sport videos using edge-based features and autoassociative neural network models. Signal, Image and Video Processing, 4(1), 61–73.
4. Wang, L. T., et al. (2020). Intelligent sports feature recognition system based on texture feature extraction and SVM parameter selection. Journal of Intelligent & Fuzzy Systems, 39(4), 4847–4858.
5. Wang, Y. G., et al. (2021). Big data and deep learning-based video classification model for sports. Wireless Communications and Mobile Computing, 2021, Article 1140611.
6. Rahman, H. M. A., et al. (2021). An efficient sports classification technique incorporating CNN transfer learning models. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC).
7. Emran, A. M. Z. M., et al. (2024). Sports video classification using convolutional neural network (CNN) with normalization flow. In 2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS).
8. Habel, O. N., et al. (2022). CLIP-ReIdent: Contrastive training for player re-identification. In Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports.
9. Minhas, J. Y. B., et al. (2019). Shot classification of field sports videos using AlexNet convolutional neural network. Applied Sciences, 9(3), 483.
10. Joshi, B. C., et al. (2020). Robust sports image classification using InceptionV3 and neural networks. Procedia Computer Science, 167, 2374–2381.
11. Ramesh, M. K., et al. (2020). A performance analysis of pre-trained neural network and design of CNN for sports video classification. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 213–216).
12. Russo, J. K. H., et al. (2019). Classification of sports videos with combination of deep learning models and transfer learning. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1–5).
13. Zhou, E. M., et al. (2024). Improved sports image classification using deep neural network and novel tuna swarm optimization. Scientific Reports, 14(1), Article 14121.
14. Russo, J. K. H., et al. (2018). Sports classification in sequential frames using CNN and RNN. In 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT) (pp. 1–3).
15. Hou, T. Z., et al. (2022). Application of recurrent neural network in predicting athletes' sports achievement. The Journal of Supercomputing, 78(4), 5507–5525.
16. Cui, Y. (2024). An efficient approach to sports rehabilitation and outcome prediction using RNN-LSTM. Mobile Networks and Applications, 1–16.

17. Yao, H. (2024). An IoT-based injury prediction and sports rehabilitation for martial art students in colleges using RNN model. Mobile Networks and Applications, 1–18.
18. Roy, C. J., et al. (2023). Multimodal fusion transformer for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–20.
19. Li, Q., et al. (2024). A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in Biology and Medicine, Article 108635.
20. Oyelade, W. H., et al. (2024). A twin convolutional neural network with hybrid binary optimizer for multimodal breast cancer digital image classification. Scientific Reports, 14(1), Article 692.
21. Xia, S. W., et al. (2024). Language and multimodal models in sports: A survey of datasets and applications. arXiv preprint.
22. Kumari, T. S., et al. (2024). Hybrid Vision Transformer and convolutional neural network for sports video classification. In 2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC) (pp. 1–5).
23. Li, H. J., et al. (2024). Research on sports image classification method based on SE-RES-CNN model. Scientific Reports, 14(1), Article 19087.
24. Liu, X. (2024). Comparison of four convolutional neural network-based algorithms for sports image classification. In 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023) (pp. 178–186).
25. Gpiosenka. (n.d.). 100 Sports Image Classification [Dataset]. Kaggle. https://www.kaggle.com/datasets/gpiosenka/sports-classification
26. Sovitrath. (n.d.). Sports Image Dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/sovitrath/sports-image-dataset
27. Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML).
28. Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint. https://arxiv.org/abs/2010.11929
29. Özdemir, E. Y., & Özyurt, F. (2025). ElasticNet-based Vision Transformers for early detection of Parkinson's disease. Biomedical Signal Processing and Control, 101, 107198.
30. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424.

VOL. 1, NO. 2, June 2026
DOI: https://doi.org/10.66395/globeis.15
ISSN: 10.66395