CricXpert: A Hybrid Spatial Fusion Model For Enhanced Player Recognition

Nadun Senarathne*, Prasan Yapa
School of Computing, Informatics Institute of Technology, 435 Galle Rd, Colombo 03, Sri Lanka

ABSTRACT

This paper introduces a hybrid spatial fusion model for real-time player recognition in T20i cricket. The system combines ResNet50 for deep feature extraction fused with Support Vector Machine (SVM), K-Nearest Neighbors (KNN) base classifiers, and a Logistic Regression meta-classifier in a stacking ensemble. A novel, domain-specific dataset was developed to capture challenges such as occlusions, variable lighting, and distant camera views followed by an expert evaluation and validation of the dataset. The proposed model achieved 98.14% accuracy, 98% precision, and 98% recall, significantly outperforming standalone deep learning models. These results highlight the advantages of combining deep and traditional machine learning methods. The approach is both robust and resource-efficient, offering a practical solution for sports analytics and laying the foundation for future work on temporal data and transfer learning.

Keywords: T20i Cricket Analytics, Player Recognition, Hybrid Fusion Models, ResNet50, Stacking Ensemble

1. INTRODUCTION

In the fast-paced format of Twenty20 International (T20i) cricket, accurate and near real-time player recognition is essential for advanced analytics and decision-making. This task becomes especially challenging during the final overs, where outcomes hinge on rapid and precise fielding analysis. However, recognition is complicated by varying lighting, occlusion, and distant camera angles. Conventional CNN-based vision systems have shown success in generic object recognition tasks[1], but their performance degrades in complex, real-world scenarios, especially when faced with small or noisy datasets[2,3]. While advanced models like ResNet and Vision Transformers offer stronger feature extraction[4,5], they often overfit and require substantial computational resources[6]. To bridge this gap, we introduce CricXpert, a novel hybrid spatial fusion model that integrates deep learning with classical machine learning. Specifically, ResNet50 is used for feature extraction, while a stacking ensemble composed of SVM, KNN, and Logistic Regression improves classification robustness and generalization. This structure separates feature extraction from decision-making and mitigates the overfitting commonly observed in end-to-end deep models.

A custom dataset reflecting real match conditions including occlusions, lighting variations, and diverse angles was curated to train and validate the system. This dataset supports the development of a player recognition model optimized for real-time deployment in cricket analytics. Our approach builds on prior hybrid methods[7,9] but advances the field by demonstrating a multi-layered ensemble fusion pipeline tailored for cricket. Experiments confirm superior performance with 98.14% accuracy, validating the approach's effectiveness in dynamic sports environments. The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 outlines the proposed model. Section 4 presents experiments and results. Section 5 discusses limitations and future directions, and Section 6 concludes the study.

2. RELATED WORK

2.1 Traditional Deep Learning Approaches in Sports Analytics

Player recognition and tracking in sports analytics have received increased attention in recent years, particularly with advancements in computer vision and deep learning. Traditional feature extraction and classification algorithms have primarily used CNNs. CNN-based models are commonly used to recognize players in dynamic sports environments because of their ability to detect and identify spatial patterns in images[1]. However, CNNs are usually limited by their tendency to overfit on small datasets and struggle in complex, real-world scenarios such as sports fields with frequent occlusions and varied lighting conditions[2].

2.2 Advanced Deep Learning Architectures

Advanced architectures, such as ResNet50 and Vision Transformers[17,18], have enhanced the capabilities of deep learning for image recognition. ResNet's residual learning capability has shown effective in reducing gradient vanishing issues, making it ideal for feature extraction in high-dimensional images[4]. Vision Transformers have demonstrated promise in capturing intricate details in visual data, but are hindered by high computational requirements, making them unsuitable for

near real-time applications[5,6]. Despite their benefits, these models frequently overfit when applied to limited, domain-specific data, such as that used in sports analytics.

2.3 Hybrid Approaches Combining Deep Learning and Machine Learning

To address the limitations of standalone deep learning models, recent studies have focused on hybrid approaches that integrate deep feature extractors with traditional machine learning classifiers to improve robustness and generalizability. For example,[7] demonstrated that combining deep features with SVM significantly improved classification in remote sensing applications, achieving a high accuracy of 95.56%. Similarly,[9] implemented a CNN-SVM hybrid model for brain tumor classification, reporting an accuracy of 94.6%, while[4] used an ensemble method with explainable AI for ovarian cancer prediction, reaching 96.3% accuracy. These studies validate the effectiveness of hybrid pipelines in complex domains, particularly when datasets are limited or noisy.

2.4 Emerging Trends in Sports Vision

Recent research has also investigated the possibility of Vision Transformers (ViTs) for fine-grained motion analysis in sports videos, which provide better spatial-temporal attention than traditional CNNs[24]. In low-data circumstances, few-shot learning techniques such as Prototypical Networks[25] have been used to recognize athletes, allowing for robust classification with minimum labeled samples. Pose-aware architectures such as ST-GCN and TokenPose[26] highlight the efficiency of exploiting skeletal and joint movement information, pointing to a possible route for improving player modeling in future CricXpert improvements.

2.5 Novelty and Contributions of The Proposed Approach

Building on these findings, this work proposes a novel spatial recognition ensemble architecture that integrates ResNet50 as a feature extractor fused with SVM and KNN classifiers, finalized through a Logistic Regression meta-layer in a stacking ensemble and a custom annotated dataset curated from real-world T20i footage covering diverse match conditions. This architecture is tailored to handle the specific challenges of sports environments, such as occlusions, variable lighting, and similar appearances among players, while maintaining real-time efficiency.

3. METHODOLOGY

This section presents the overall design of the proposed hybrid spatial fusion recognition system shown in Figure 1. This framework outlines the major steps, from data acquisition and preprocessing to model training and player classification, while providing a high-level overview of the system's structure and workflow. The goal was to develop a robust player recognition model that successfully addresses overfitting and performs consistently in dynamic cricket environments.

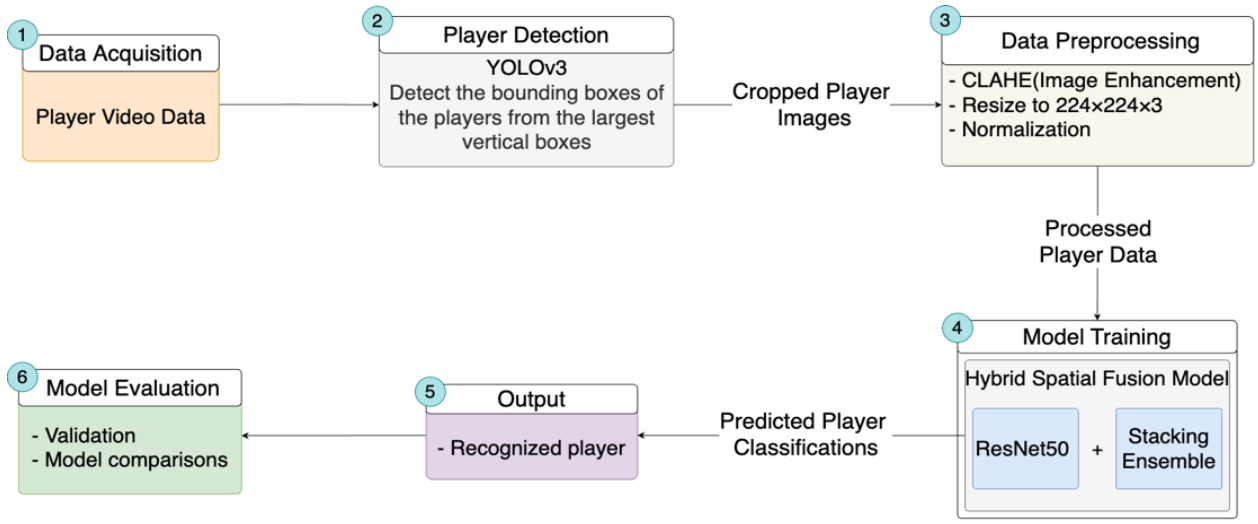

Figure 1: High Level Design of the Hybrid Spatial Fusion Recognition System for T20i Cricket Player Recognition.

3.1 Data Pre-processing

Player detection accuracy directly impacts the recognition performance. To address variability in player visibility and image quality, the YOLOv3[23] model is employed solely for initial player detection in each video frame. Since YOLO often identifies multiple bounding boxes per frame, the largest vertical box is selected, corresponding to cricket players' typical upright posture. Extracted player images then get cropped and undergo Contrast Limited Adaptive Histogram Equalization

(CLAHE) for explicit image enhancement, addressing challenges posed by varying lighting conditions and occlusions, ensuring high-quality feature extraction.

3.2 Feature Extraction

Enhanced images are processed using the pretrained ResNet50 model[17], chosen due to its robust residual learning capability that effectively mitigates issues such as vanishing gradients. Preprocessed images are resized to 224×224×3 and passed through a pretrained ResNet50, with the top classification layers removed. The 2048-dimensional output from the final pooling layer serves as a compact and robust spatial feature vector for classification. These vectors are used as input to the base classifiers in the stacking ensemble.

3.3 Stacking Ensemble for Classification

To improve classification robustness, a stacking ensemble approach combines traditional machine learning classifiers SVM[19], KNN[20] with Logistic Regression[21] as a meta-classifier (Figure 2).

Workflow of the Stacking Ensemble

Base Classifiers:

- **SVM:** Utilizes an RBF (Radial Basis Function) kernel to distinguish non-linear patterns within the feature space derived from ResNet50. SVM reduces noise and improves the model's resilience to data fluctuations.
- **KNN:** Captures local spatial relationships in the feature space using the Manhattan distance metric. The ability for effective generalization enhances the accuracy of SVM, establishing a balanced and efficient basis for classification.

Meta-Classifier:

- **Logistic Regression:** Functions as the ultimate decision-making tier, integrating predictions from SVM and KNN. The meta-classifier combines these predictions into a unified output, minimizing individual classifier inaccuracies and enhancing overall accuracy. By learning patterns in the outputs of the base classifiers, Logistic Regression ensures that the ensemble effectively addresses variability in cricket player data.

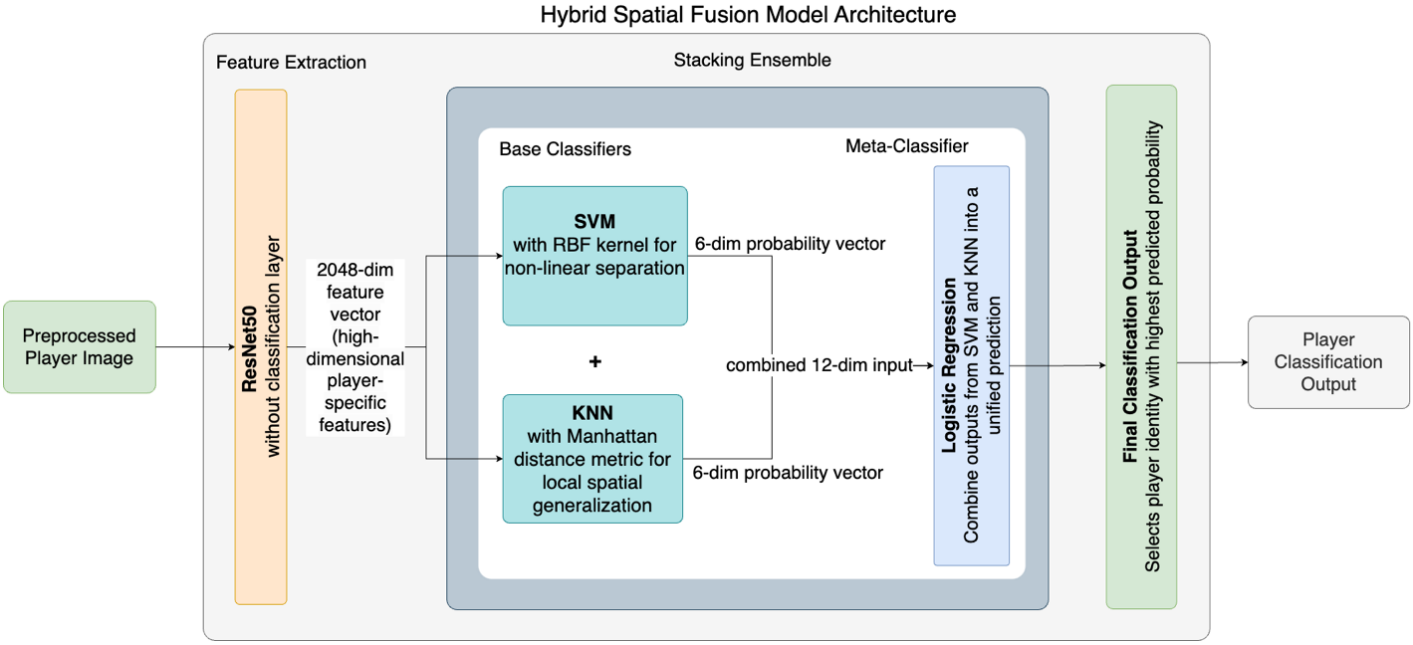

Figure 2: Detailed Hybrid Spatial Fusion Model Architecture for the Proposed T20i Cricket Player Recognition System.

Probability outputs from these base classifiers serve as inputs to a Logistic Regression meta-classifier, combining predictions into a unified and robust final decision, reducing classification errors and significantly enhancing overall accuracy. Each base classifier (SVM and KNN) produces a probability distribution across the six player classes, resulting in two separate 6-dimensional output vectors. These are concatenated to form a single 12-dimensional feature vector, which serves as the input to the Logistic Regression meta-classifier. This meta-layer then outputs the final predicted player class by learning from the combined decision patterns of the base classifiers.

Improvement Over Traditional CNN + SVM/KNN Approaches:

Unlike traditional pipelines where a CNN is followed by a single classifier (such as SVM or KNN), our approach employs a stacking ensemble that integrates multiple classifiers. This layered structure allows each base classifier to specialize in capturing different aspects of the feature space, SVM excels at handling non-linear decision boundaries, while KNN captures local relationships. The Logistic Regression meta-classifier learns from the strengths and weaknesses of both, leading to improved robustness and reduced overfitting. This multi-stage learning strategy significantly outperforms single-classifier CNN-based methods, as demonstrated in the below Section 4 through empirical results and statistical validation.

3.4 Hyperparameter Optimization

Tuned hyperparameters and cross-validation significantly reduced overfitting, resulting in a model that effectively tackles the specific issues of T20i cricket player recognition. The SVM used an RBF kernel with a regularization parameter of C = 1 and a gamma value set to 'scale' to balance bias and variance. The KNN classifier was configured with k = 3 neighbors, employing the Manhattan distance metric and uniform weighting to enhance local spatial relationships in the feature space. The meta-classifier was trained using Logistic Regression with a regularization parameter of C = 0.001, an L2 penalty for regularization, and the 'liblinear' solver for optimization.

4. EXPERIMENTS

This section presents the proposed spatial fusion model's evaluation results, which include comparisons of performance and robustness across several architectures. The model's ability to recognize T20i cricket players was evaluated using three standard metrics: accuracy, which reflects the overall proportion of correctly identified players; precision, which quantifies the proportion of correctly predicted player instances among all positive predictions; and recall, which assesses the model's ability to identify actual players present in the frames. These metrics collectively provide a comprehensive evaluation of both the correctness and completeness of the recognition system.

4.1 Dataset

To address the lack of publicly available datasets for cricket player recognition under dynamic match situations, a novel dataset was curated expressly for the proposed method using frames extracted from high-resolution T20i match broadcasts publicly available on official YouTube channels of international cricket boards. The dataset includes annotated images of six cricket players, each with 130-150 images taken from various vantage points and match conditions. Frames were extracted using YOLOv3, from match recordings at a consistent sampling rate to balance image diversity and homogeneity and manually verified for label correctness. Figure 3 illustrates representative samples from the dataset used in the evaluation.

Figure 3: Sample annotated player images from the CricXpert dataset.

Expert Evaluation and Validation of Dataset Quality

To assess the dataset's quality, defined by **annotation accuracy** and **diversity of visual conditions** and its **relevance to real-world T20 match scenarios**, a structured expert validation was conducted. Three independent professionals (Subjects 1–3) with backgrounds in cricket coaching, analytics, and applied data science were selected. Inclusion criteria required experience in sports analytics, availability for a live demonstration, and completion of a structured evaluation form. Individuals without cricket domain expertise or who declined formal feedback were excluded. Each expert participated in a live demo session, followed by a standardized Google Form questionnaire and rated the key abovementioned dataset attributes on a 5-point Likert scale (1 = Poor, 5 = Excellent). Since no personal or sensitive data were collected, ethical approval from a review board was not required.(Figure 4)

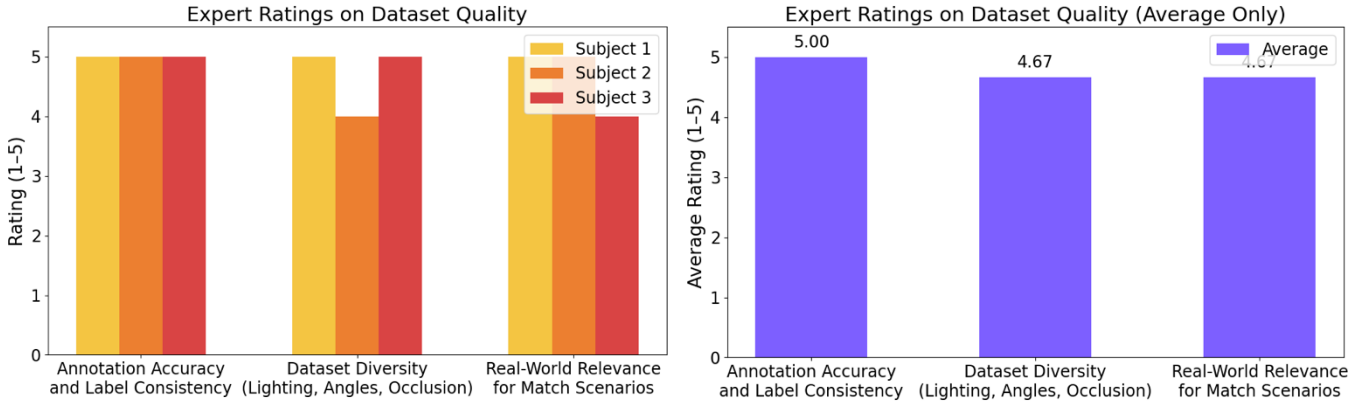

Figure 4: Expert evaluation of dataset quality based on the three key criteria.

4.2 Model Comparison and Performance Metrics

The initial stage of model selection involved examining various deep learning architectures, starting with a customized CNN model with six convolutional blocks, which were then followed by batch normalization, ReLU activation, and max pooling. This was followed by five dense layers with dropout for regularization, before the final SoftMax classification layer. However, this approach resulted in poor performance, highlighting the need for more advanced architectures. Several models were then assessed and among these, ResNet50 emerged as the best performer due to its robust residual learning capability, which effectively addressed vanishing gradient issues and allowed for deeper feature extraction. The initial model comparisons yielded the following results as shown by below Table 1.

Table 1.Baseline Model Performance Metrics

Model	Accuracy(%)	Precision(%)	Recall(%)
Custom CNN	40.32	41	40
DenseNet[10]	69	70	68
EfficientNetB0 [11]	57.82	57	56
Inception [12]	44	46	45
MobileNetV2 [13]	61.22	61	61
VGG16 [14]	60.78	61	60
NASNet [15]	57	59	57
Xception [16]	59	57	58
Vision Transformers (ViT) with different configurations	~42-60	~43-60	~42-60
ResNet50 [17]	61.75	62	61

Note: All baseline models were re-implemented and fine-tuned on our custom dataset under identical preprocessing and training conditions for a fair comparison. (ResNet50 was utilized as a frozen feature extractor in the final model.)

4.3 Comparison with Vision Transformers

To further explore potential improvements, Vision Transformers (ViTs) were evaluated on the dataset, with five distinct ViT models trained over various epochs and different configurations. Despite the promise exhibited in other domains, ViTs proved computationally expensive, with a single epoch taking over an hour and forty minutes. Furthermore, they exhibited a strong tendency to overfit, achieving only little improvements in accuracy over ResNet50 but at a large computational expense. This demonstrated that ResNet50 was better suited for near real-time, resource-efficient player recognition in sports environments.

4.4 Deep Feature Fusion with ML Classifiers

While ResNet50 outperformed earlier models, overfitting persisted despite early stopping and parameter adjustments. This led to the hypothesis that these architectures' classification layers were responsible for the overfitting. To address this, a

fusion technique was implemented: ResNet50 was used solely for feature extraction. Initial experiments with a few ML classifiers revealed that SVM and KNN performed well, enhancing classification accuracy while reducing overfitting. The following Table 2 summarize the improvement:

Table 2. Performance Metrics of Resnet50 Fused with Different ML Classifiers

Model	Accuracy(%)	Precision(%)	Recall(%)
ResNet50 + SVM	95.78	96	96
ResNet50 + KNN	95.43	94	95
ResNet50 + Random Forest	94	93	94
ResNet50 + Gradient Boost Machine	90.36	91	90
ResNet50 + Decision Tree	66.22	66	66

4.5 Final Stacking Ensemble Results

The final stacking ensemble technique, which combined ResNet50, SVM, and KNN with a final Logistic Regression classification layer, demonstrated the highest robustness and accuracy among all configurations tested. The stacking ensemble efficiently addressed overfitting by exploiting the capabilities of many classifiers while preserving excellent generalizability across a wide range of conditions, including low-light and occluded player scenarios. Several ensemble strategies were investigated, including Voting Classifiers and Decision-Level Fusion, however the stacking ensemble method outperformed them all, displaying higher accuracy and robustness. Table 3 summarizes the ensemble approaches' outcomes.

Table 3. Performance Metrics of Ensemble Techniques

Model	Accuracy(%)	Precision(%)	Recall(%)
Voting Classifier	95	95	95
Decision-Level Fusion	96.27	96	96
Stacking Ensemble	98.14	98	98

Performance Visualization: To validate the stacking ensemble's performance, the results were compared to the baseline ResNet50 model using learning curves and confusion matrices.

1. Learning Curve Comparison:

Baseline Model (ResNet50): The training vs. validation loss and accuracy curves for ResNet50 (Fig. 5, left side) highlight noticeable overfitting. The training accuracy rapidly increases over epochs, but the validation accuracy plateaus, indicating limited generalizability. Similarly, while the loss curves converge, the gap between training and validation loss remains visible, further suggesting overfitting in complex cricket scenarios.

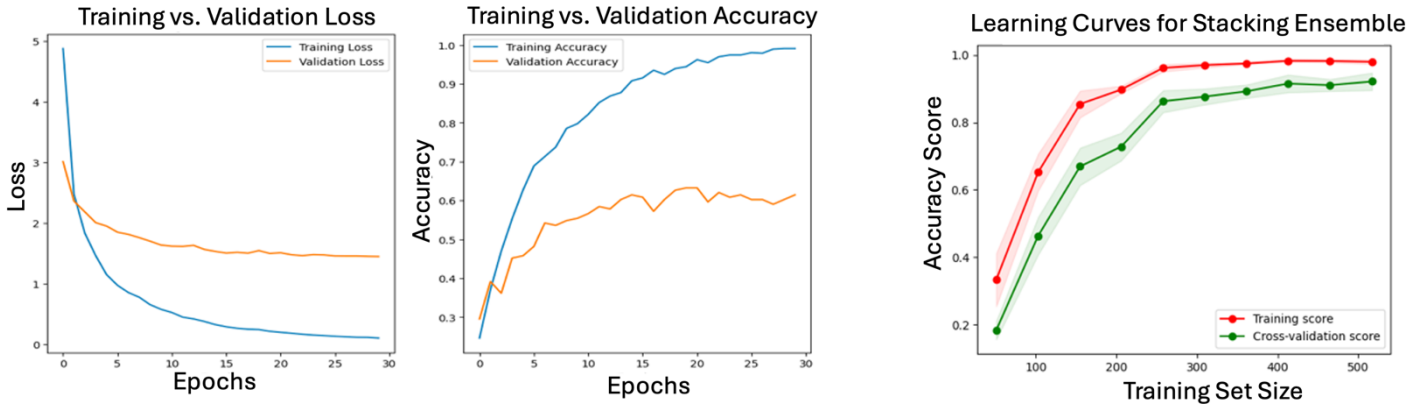

Figure 5: Training vs. Validation Loss and Accuracy for Baseline ResNet50 Model (Left) and Learning Curve for Stacking Ensemble Method (Right).

Stacking Ensemble: The learning curve for the stacking ensemble (Fig. 5, right side) demonstrates the mitigation of the overfitting issue, with training and cross-validation accuracy scores converging as training size increases. This demonstrates the ensemble's capacity to generalize efficiently in an array of situations, including low light and occluded environments.

2. Confusion Matrix Comparison:

Baseline Model (ResNet50): The confusion matrix for ResNet50 (Fig. 6, left side) shows multiple misclassifications, notably for players with similar jersey numbers or who are obscured by other players. Higher False Positives (FP) and False Negatives (FN) show that the basic model struggled with complicated scenarios.

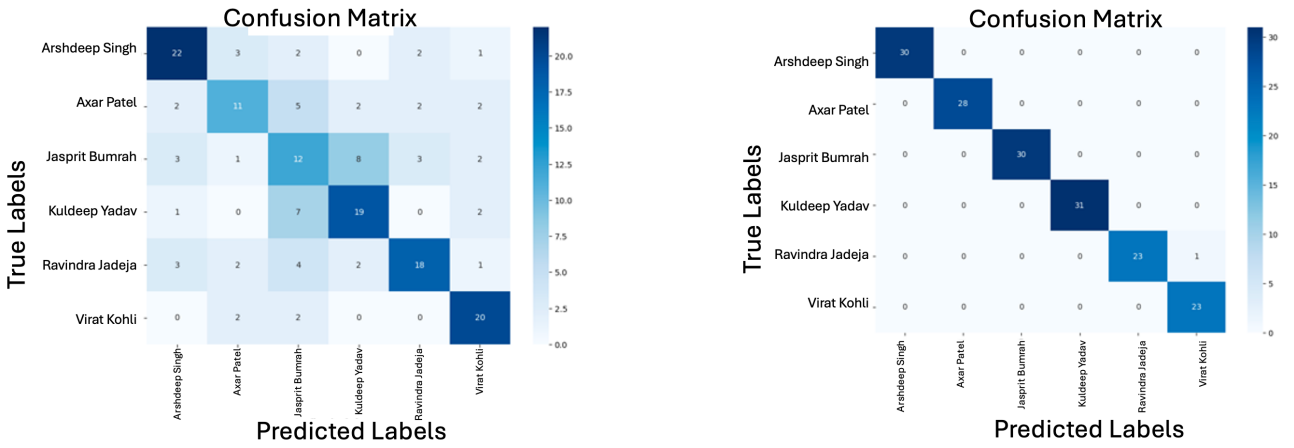

Fig. 6. Confusion Matrix for Baseline ResNet50 Model (Left) and for Stacking Ensemble Model (Right).

Stacking Ensemble: The confusion matrix for the stacking ensemble (Fig. 6, right side) demonstrates substantial improvement, with higher True Positives (TP) and significantly reduced FP and FN values. This improvement is the direct result of the ensemble's superior and robust classification capabilities.

These findings indicate that the stacking ensemble method, with tailored hyperparameters and cross-validation, outperformed the other methods, reducing overfitting significantly and providing reliable player recognition in real-world circumstances. While formal statistical significance testing was not conducted, the stacking ensemble consistently showed performance increases across several cross-validation folds. These advances are further corroborated by learning curve and confusion matrix analysis, which demonstrate less overfitting and better generalization.

Reproducibility, Code and Dataset Availability

The full codebase, along with preprocessing scripts and configuration files, is available at https://github.com/Nadun999/FYP to facilitate reproducibility. In addition, the annotated cricket player dataset utilized in this work, which includes samples from real T20i match situations, will also be released in the same repository. This release is intended to facilitate replication, benchmarking, and further innovation in cricket-specific computer vision applications.

5. DISCUSSION

Findings of the proposed method demonstrate that combining deep learning feature extraction with machine learning classifiers produces significant advantages for player recognition in the complex and dynamic environment of T20i cricket. The spatial model effectively mitigates overfitting issues encountered in standalone deep learning models by employing ResNet50 for feature extraction as well as SVM, KNN, and a final Logistic Regression layer in a stacking ensemble. This hybrid approach leverages the strengths of each classifier, by achieving high accuracy and generalizability across variable conditions, such as low-light and occlusions.

5.1 Benefits of the Hybrid Approach

The hybrid model, which included ResNet50 and machine learning classifiers, showed significant improvements in performance metrics, particularly in accuracy and robustness. This approach helps overcome one of the main limitations of deep learning models, overfitting on small or domain-specific datasets, by employing simpler classifiers that generalize well without the need for extensive computational resources. The stacking ensemble method further enhances these

benefits by combining the predictions of multiple classifiers, allowing the model to correct errors that individual classifiers may produce, especially in complex settings where limited training data is available.

Unlike traditional CNN+SVM or ML combinations that use end-to-end deep classification, our method explicitly separates feature learning from classification, and our stacking ensemble combines multiple classifiers in a layered structure, achieving higher robustness and generalization in challenging real-world cricket scenarios. A standard M1 MacBook Pro without GPU acceleration was used to deploy and test the final model. The model can be used for near-real-time sports analytics applications as the average inference time for identifying a player from a video clip was between 2 and 5 seconds. While emerging efficient fine-tuning methods such as LoRA may reduce ViT training time, they were not explored in this study due to resource constraints. This level of efficiency is a direct result of using a frozen ResNet50 feature extractor combined with lightweight machine learning classifiers in the classification stage. The proposed ensemble model achieves a better trade-off between performance, computational efficiency, and practical deployability.

5.2 Limitations and Trade-offs

While the stacking ensemble approach produced positive developments by mitigating the overfitting issue, some limitations, such as the complexity introduced to model deployment via the integration of multiple classifiers, are still observed. Furthermore, while cross-validation and hyperparameter tuning improved model robustness, the ensemble approach may still require additional tuning to adapt to different sports or environmental variables, such as varying field sizes or camera angles specific to each venue. The dataset used in this study was designed and annotated specifically for this method. Lastly, while the hybrid model produced robust results, there are instances with considerable occlusion or overlapping players still presenting challenges.

6. CONCLUSION

This paper introduced a hybrid spatial fusion model for real-time player recognition in T20i cricket, combining ResNet50 feature extraction with a stacking ensemble fused of SVM, KNN, and Logistic Regression. This architecture addresses key challenges in sports analytics, including overfitting, variable lighting, occlusions, and limited training data. Experimental findings demonstrate that the stacking ensemble method effectively enhances model generalizability and performance, while achieving high accuracy in complex cricket scenarios. By fusing deep learning and machine learning techniques, this study contributes a practical, resource-efficient solution for real-time player recognition in sports analytics. Findings of this study, indicate that the proposed model not only meets the needs of T20i cricket, but also holds potential for broader applications across other sports and dynamic environments.

Future work will explore integrating additional data modalities, such as temporal movement patterns, or exploring more advanced ensemble strategies, to complement the spatial recognition model and to further enhance accuracy in scenarios with significant occlusion. Additionally, applying transfer learning techniques to adapt the ResNet50 backbone for new sports datasets can extend the model's applicability without extensive retraining. While the base components (ResNet50, SVM, KNN) are well-established, the novelty lies in their integration for cricket-specific recognition and real-time analytics, a space with very limited prior work. To further enhance performance and portability, we intend to investigate adaptive ensemble weighting techniques and lightweight ResNet adaptations in the future. To increase the model's applicability, we intend to use transfer learning to extend it for comparable sports like baseball and soccer by utilizing similar posture and movement patterns. We also intend to concentrate on expanding the dataset to include more players, matches, and venues. This study lays the groundwork for robust, real-time player recognition systems, advancing the capabilities of sports analytics and computer vision in high-performance domains.

REFERENCES

[1] X. Han, Y. Zhang, M. Liu, and Z. Wang, "A robust and consistent stack generalized ensemble-learning framework for image segmentation," *Journal of Engineering and Applied Science*, 2023.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] M. Gao, J. Li, and L. Zhao, "Exploring the combination of CNN and transformer models for multi-modal image analysis," in *Proceedings of the 2022 International Conference on Machine Learning and Applications*, 2022.

[3] Y. Wu, Y. He, and Y. Wang, "Multi-class weed recognition using hybrid CNN-SVM classifier," *Sensors*, vol. 23, no. 16, p. 7153, 2023.

[4] M. Shaikh, F. Alsunaidi, and S. Alamoudi, "Improved prediction of ovarian cancer using ensemble classifier and Shaply explainable AI," *MDPI*, 2022

[5] S. Guha, A. Kumar, and S. Dey, "Explainable AI for interpretation of ovarian tumor classification using enhanced ResNet50," *MDPI*, 2024

[6] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding Robustness of Transformers for Image Classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 10211–10221.

[7] F. Özyurt, "Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures," *The Journal of Supercomputing*, vol. 76, pp. 1–19, 2020.

[8] H. Kibriya, M. Rahman, R. Ferdous, and S. Mahmud, "A novel and effective brain tumor classification model using deep feature fusion and famous machine learning classifiers," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7897669, 2022.

[9] H. Kibriya, M. Rahman, R. Ferdous, and S. Mahmud, "Multiclass brain tumor classification using convolutional neural network and support vector machine," in *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*, Karachi, Pakistan, 2021, pp. 1–4.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269.

[11] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Sep. 11, 2020, *arXiv*: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.

[12] C. Szegedy *et al.*, "Going Deeper with Convolutions," Sep. 17, 2014, *arXiv*: arXiv:1409.4842. doi: 10.48550/arXiv.1409.4842.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Mar. 21, 2019, *arXiv*: arXiv:1801.04381. doi: 10.48550/arXiv.1801.04381.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[15] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," Apr. 11, 2018, *arXiv*: arXiv:1707.07012. doi: 10.48550/arXiv.1707.07012.

[16] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," Apr. 04, 2017, *arXiv*: arXiv:1610.02357.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[19] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[20] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[21] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958.

[22] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[23] J. Redmon and A. Farhadi, "YOLOv3: An incremental Improvements," arXiv:1804.02767,2018

[24] A. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?", in ICML, 2021.

[25] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning", in NeurIPS, 2017.

[26] Y. Zhang et al., "ST-GCN: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", in AAAI, 2018