CricXpert: A Hybrid Spatial Fusion Model For Enhanced Player Recognition

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country email
address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country email
address or ORCID

Abstract—**In the realm of Twenty20 International (T20i) cricket, real-time, precise player recognition is critical to enhancing game analytics and strategic decision making. This study presents a novel hybrid spatial fusion recognition model that synergistically integrates ResNet50 for feature extraction with machine learning classifiers such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and a final Logistic Regression layer within a stacking ensemble framework. To address the challenges provided by variable lighting, occlusions, and distant camera angles, a novel dataset, that reflects the dynamic conditions inherent in T20i cricket, was curated. Through meticulous hyperparameter tuning and rigorous cross-validation, this model achieved a commendable accuracy of 98.14%, precision of 98%, and recall of 98%, outperforming standalone deep learning architectures. This study demonstrates the efficacy of ensemble techniques in complicated, real-world recognition tasks, paving the path for future research into temporal data integration and transfer learning to enhance model adaptability and performance across a wide range of sports analytics applications.**

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

I. Introduction

In the fast-paced and highly competitive domain of Twenty20 International (T20i) cricket, accurate and quick player recognition is critical for enhancing both game analytics and viewer experience. This is especially true in the last few overs of a match, where critical decisions are made and fielders' performance can have a substantial impact on the match outcomes. However, recognizing players in such dynamic environments raises a number of challenges. Variable lighting conditions, player-caused occlusions, and distant camera angles are few of such causes identified challenges that hinders the recognition process. Traditional computer vision techniques, such as Convolutional Neural Networks (CNNs), have advanced significantly in recognizing objects and features in images [1]. However, their effectiveness in complicated real-world settings, such as sports grounds, remains inconsistent, especially when working with smaller datasets or noisy data [2], [3].

Deep learning models such as ResNet and Vision Transformers have been employed, in the context of cricket player recognition, for spatial feature extraction due to their success in capturing complex image patterns [4], [5]. Despite their power, where the variability of the data is high, these models are prone to overfitting and lack the robustness needed for real-world sports analytics. [6]. A significant obstacle in this domain has been the scarcity of publicly available datasets that are tailored for cricket player recognition, specially under dynamic match settings. To bridge this gap, a novel dataset was created specifically to support the proposed method. This dataset contains labelled player images captured across diverse match scenarios, including variations in lighting, occlusions, and various camera angles, and it provides a robust foundation for training and evaluating the proposed spatial recognition model, thereby ensuring relevance to the unique challenges of T20i cricket analytics.

In order to enhance model performance, recent studies have explored hybrid models that combine deep learning feature extractors with traditional machine learning classifiers, such as Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) [7], [8]. For an example, the deep feature selection technique proposed by Özyurt (2020) using fused deep learning architectures, highlighted the advantages of integrating deep learning with machine learning classifiers to improve performance in complex environments [7]. Furthermore, Kibriya et al. (2021) explored similar hybrid approaches for brain tumour classification using CNN-SVM models, which demonstrated the effectiveness of such combinations for medical image classification tasks [8], [9].

The proposed method addresses these challenges by enhancing the performance of spatial recognition models through a novel fusion of deep learning feature extractors with traditional machine learning classifiers. We utilize ResNet50 as the primary feature extractor and combine it with SVM and KNN classifiers and a final layer of Logistic Regression in a stacking ensemble. The key point in employing stacking ensembles is due to the improved model accuracy that has been shown and the robustness achieved by leveraging the strengths of different models to correct each other's errors [4], [5]. By integrating machine learning classifiers with deep learning models, particularly under challenging conditions such as low-light environments and occlusions, we mitigate overfitting and improve generalizability.

The subsequent part of the paper is organized as follows: Section II reviews the existing literature and identifies the gaps addressed by this work, with a focus on hybrid models and advanced deep learning architectures. Section III details the approach for the proposed hybrid spatial fusion model, which includes data preprocessing, model architecture, and ensemble techniques. Section IV presents the experimental setup, evaluation metrics, and findings of the model, emphasizing its performance under challenging conditions. Section V discusses the study's implications and limitations. Finally, Section VI concludes the study by summarizing the contributions, future directions and potential applications of the proposed approach.

II. Related Work

A. 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].

B. 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 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.

C. Hybrid Approaches Combining Deep Learning and Machine Learning

For addressing the constraints of standalone deep learning models, recent research has focused on hybrid methods that integrate deep learning feature extractors with machine learning classifiers to improve model robustness and generalizability. Özyurt (2020) observed that combining deep learning models with SVM [19] and KNN [20] classifiers can mitigate overfitting and enhance classification performance in demanding environments [7]. Similarly, Kibriya et al. (2021) employed a CNN-SVM hybrid model to classify brain tumors, illustrating that combining a deep feature extractor and a machine learning classifier can increase model accuracy while reducing processing demands in medical imaging tasks [8], [9].

D. Contribution of The Proposed Approach

This work extends these findings by creating a spatial recognition model that combines ResNet50 and machine learning classifiers. Our approach addresses common overfitting issues in deep learning models employing a stacking ensemble of SVM and KNN classifiers, as well as a final Logistic Regression layer, to provide an efficient, real-time solution for recognizing cricket players in variable and challenging field conditions.

III. Methodology

This section presents the overall design of the proposed hybrid spatial fusion recognition system shown in Fig. 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.

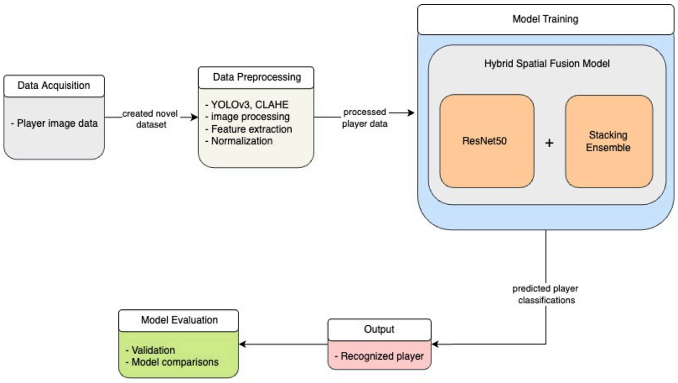


Fig. 1. Overall Design of the Hybrid Spatial Fusion Recognition System for T20i Cricket Player Recognition.

Player recognition in sports analytics, particularly in the fast-paced context of T20i cricket, poses unique obstacles that limit the accuracy and consistency of typical recognition models. Real-time identification of fielders during the last few overs of a T20i match is crucial since it has a direct impact on both strategic decisions and statistical analyses. However, varying lighting conditions, frequent occlusions caused by overlapping players, and distant camera angles make accurate recognition extremely difficult. Traditional computer vision and deep learning models, such as CNNs and Vision Transformers [18], perform well in object and pattern recognition but fall short in highly variable environments. These models are especially prone to overfitting on limited training data, decreasing their capacity to generalize in real-world scenarios. Furthermore, the processing requirements for high-performance models, such as Vision Transformers, render them unsuitable for real-time applications.

To address these issues, this study develops a spatial recognition model that integrates ResNet50 [17] for feature extraction with machine learning classifiers including SVM, KNN, and a final layer of Logistic Regression [21] within a stacked ensemble [22] model for classification. This hybrid strategy takes advantage of deep learning's feature extraction power while improving resilience and reducing overfitting with simpler machine learning classifiers, resulting in improved accuracy and efficiency under demanding environments.

A. Data Pre-processing

To accurately separate players from the background, the YOLOv3 model was employed for player detection (Fig. 1). The data preprocessing phase includes YOLOv3 for player detection and CLAHE (Contrast Limited Adaptive Histogram Equalization) for image enhancement, ensuring that only high-quality, focused player images are inputted into the feature extraction model. YOLOv3's ability to detect objects in real time with great precision makes it ideal for segmenting players from video frames. However, because numerous boxes are frequently identified in each frame, an extra selection criterion was used: the largest vertical bounding box was chosen, as the player's upright stance while running or walking makes them the most visible vertical object in the frame. CLAHE was also utilized to improve image quality before ResNet50 was used to extract features (Fig. 1). This method reduced background noise while focusing detection on the player.

B. Model Selection and Evaluation

The initial stage of model selection involved evaluating different deep learning architectures, beginning with a custom CNN model. However, this approach resulted in poor performance, highlighting the need for more advanced architectures. Several models were then assessed, including DenseNet [10], EfficientNet [11], Inception [12], MobileNet [13], VGG16 [14], NASNet [15], Xception [16], and ResNet50. 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.

To further explore advanced architectures, Vision Transformers (ViTs) were examined using five different ViT models spanning several epochs and configurations. Despite their potential, ViTs proved to be computationally expensive, necessitating extensive training time of around 100 minutes per every epoch. Furthermore, they were prone to overfitting on the limited dataset given, rendering them inappropriate for real-time recognition in this scenario. As a result, ResNet50 was chosen as the principal feature extraction model due to its balance of performance and computational efficiency.

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, while SVM and KNN were used for classification. The extracted features were classified using a stacking ensemble of SVM and KNN, with a final Logistic Regression classification layer. Hyperparameter tuning was carried out on all components, resulting in effective overfitting mitigation and improved model performance.

Fig. 2 illustrates the detailed architecture of the hybrid spatial fusion recognition model. This diagram depicts the individual components and interactions of the feature extraction layer (ResNet50), base classifiers (SVM and KNN), and meta-classifier (Logistic Regression). Each stage in the architecture is intended to reduce overfitting while maintaining robustness in demanding conditions

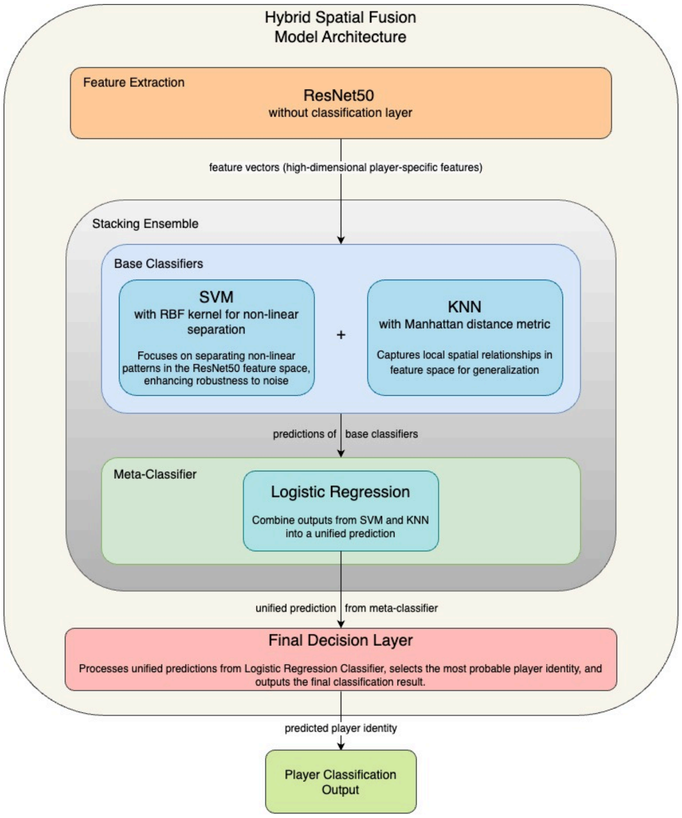

Hybrid Spatial Fusion
Model Architecture
Feature Extraction
ResNet50
without classification layer
feature vectors (high-dimensional player-specific features)
Stacking Ensemble
Base Classifiers
SVM
with RBF kernel for non-linear separation
Focuses on separating non-linear patterns in the ResNet50 feature space, enhancing robustness to noise
+
KNN
with Manhattan distance metric
Captures local spatial relationships in feature space for generalization
predictions of base classifiers
Meta-Classifier
Logistic Regression
Combine outputs from SVM and KNN into a unified prediction
unified prediction from meta-classifier
Final Decision Layer
Processes unified predictions from Logistic Regression Classifier, selects the most probable player identity, and outputs the final classification result.
predicted player identity
Player Classification Output

Fig. 2. Detailed Hybrid Spatial Fusion Model Architecture for T20i Cricket Player Recognition.

C. Stacking Ensemble for Classification

To enhance classification performance, multiple machine learning classifiers were tested, including the SVM, KNN, Random Forest, Gradient Boost Machine, and Decision Tree classifiers. SVM and KNN outperformed the other models in terms of overfitting and classification accuracy. However, to further improve robustness, a stacking ensemble method was introduced.

The proposed approach employs a stacking ensemble architecture, which utilizes the strengths of machine learning classifiers to supplement the deep learning feature extractor. The stacking ensemble combines SVM and KNN as base classifiers, with Logistic Regression as the meta-classifier. The stacking ensemble improves robustness and generalizability by combining the outputs of many classifiers, thereby minimizing overfitting and enhancing performance under challenging conditions.

Features were extracted from pre-processed player images using ResNet50, a deep learning architecture known for its residual learning ability. ResNet50 was utilized with its classification layer removed to ensure that the retrieved feature vectors were optimal for classification. These features were then normalized and underwent dimensionality reduction to ensure compatibility with the subsequent machine learning classifiers. Fig. 2 elaborates the flow of features from ResNet50 to the stacking ensemble, demonstrating how the base classifiers (SVM and KNN) collaborate with the meta-classifier (Logistic Regression) to produce a unified prediction.

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.

Multiple ensemble techniques, such as Voting Classifiers and Decision-Level Fusion, were evaluated. The stacking ensemble method surpassed conventional techniques in both accuracy and robustness. The stacking ensemble, which integrates ResNet50 for feature extraction, SVM and KNN for base classification, and Logistic Regression as the meta-classifier, attained exceptional performance in challenging conditions, including low-light environments, occlusions, and distant camera angles.

Tuned hyperparameters and cross-validation significantly reduced overfitting, resulting in a model that effectively tackles the specific issues of T20i cricket player recognition. For the stacking ensemble, 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. These hyperparameters were determined through cross-validation to ensure optimal performance. This stacking ensemble method improves classification accuracy while offering a computationally efficient and scalable solution, indicating its potential for broader applications in sports analytics.

IV. Experiments

In this section, we present the results from the evaluations of the spatial fusion model, comparing several architectures and configurations to evaluate the performance and resilience of the final solution. The performance of the spatial recognition models for T20i cricket player recognition was assessed using the following metrics:

1. Accuracy: Accuracy assesses the model's overall correctness by comparing the quantity of accurately identified players to the total number of players evaluated. This is especially beneficial for evaluating overall model efficacy in situations where positive and negative predictions are of equal importance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision: Precision assesses the accuracy of the model by determining the proportion of correctly predicted players. In player recognition, high precision minimizes erroneous detections, such as background objects that can be misidentified as players.

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. Recall: Recall measures the model's completeness by determining the proportion of actual players in the frames that were accurately identified. It is essential in cricket analytics to guarantee that no player goes unrecognized during high-stakes moments.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where:

True Positives (TP): Correctly identified players.
True Negatives (TN): Correctly rejected non-players.
False Positives (FP): Incorrectly identified as players.
False Negatives (FN): Players that the model failed to recognize

A. Dataset

To address the lack of publicly available datasets for cricket player recognition under dynamic match situations, a novel dataset was created expressly for the proposed method. The dataset includes annotated images of six cricket players, each with 120-150 images taken from various vantage points and match conditions.

The data collecting process included selecting high-resolution footage from T20i matches to ensure a diverse representation of camera angles, lighting conditions, occlusions, and player movement. Frames were extracted from match recordings at a consistent sampling rate to balance image diversity and homogeneity. Preprocessing steps included converting raw frames to standard image dimensions, detecting bounding boxes to segregate players with YOLOv3, and improving image quality with CLAHE. To validate the dataset, Large Language Model (LLM) validation was utilized, which takes advantage of cutting-edge natural language understanding and pattern recognition technologies. The LLM evaluated annotations against predefined validation criteria to ensure accuracy, consistency, and relevance for the intended task. This approach provides a scalable and objective means of verifying dataset quality, resulting in accurate player recognition.

Validation of the Dataset and Results

The credibility of the novel cricket player recognition dataset was established via LLM-based validation utilizing the pretrained CLIP model. The validation approach includes testing the model's capacity to distinguish between six cricket players using embeddings created for each image in the dataset. This was performed under diverse conditions without modifying the dataset, ensuring its integrity.

1. Overall Accuracy: The model attained an accuracy of 82.00% after adding a confidence threshold, demonstrating the dataset's durability in supporting recognition tasks in dynamic environments. Predictions with confidence scores below the threshold were labeled as "uncertain," which reduced misclassification.

2. LLM Evaluation Graph: A prediction distribution graph (Fig. 3) was created to assess the model's output across all categories. The majority of uncertain predictions were for challenging classes (Player 1 and Player 4), demonstrating the efficiency of the thresholding technique.

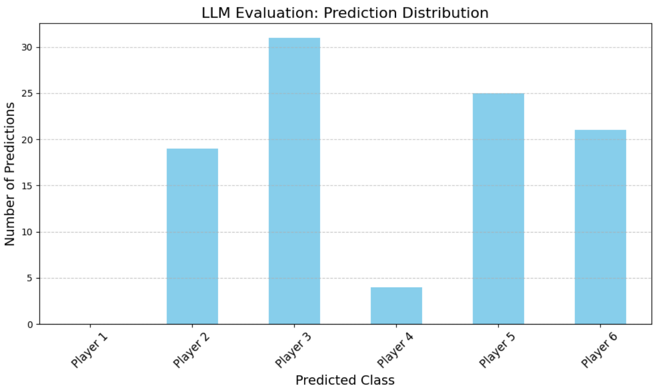


Fig. 3. LLM Evaluation grpah of the Prediction Distribution.

These findings support the dataset's credibility as a dependable resource for cricket player recognition tasks in real-world settings. Its validation confirms its viability for usage in this project and future research in sports analytics. However, minor deviations in predictions for challenging classes (e.g., Player 1 and Player 4) were observed during the error analysis, indicating areas for further refinement. These deviations, analyzed in the Discussion section, highlight the robustness of the proposed method in addressing such errors and maintaining overall reliability.

B. Model Comparison and Performance Metrics

To determine the most successful model for player recognition, we initially compared different deep learning architectures, including a custom CNN, DenseNet, EfficientNet, Inception, MobileNet, VGG16, Xception, NASNet, and ResNet50. Each model was tested for accuracy, precision, and recall, and ResNet50 emerged as the best performer across all metrics. The initial model comparisons yielded the following results:

- **Custom CNN**: It demonstrated poor accuracy and excessive overfitting, making it inappropriate for complicated cricket environments.
- **DenseNet, EfficientNet, Inception, MobileNet, VGG16, NASNet, Xception**: These models performed reasonably well, but exhibited limited generalizability, especially in low-light and obstructed conditions.
- **ResNet50**: Superior feature extraction capabilities and robustness were demonstrated, resulting in higher accuracy with lower overfitting than other deep learning models.

TABLE I
BASELINE MODEL PERFORMANCE METRICS

Model	Accuracy(%)	Precision(%)	Recall(%)
Custom CNN	30	31	30
DenseNet	69	70	68
EfficientNetB0	92.82	93	92
Inception	44	46	45
MobileNetV2	71.82	74	71
VGG16	84	85	84
NASNet	57	59	57
Xception	61	62	61
ResNet50 (20 epochs)	88	89	88
ResNet50 (30 epochs)	95.18	96	95
Vision Transformers (ViT) with different configurations	~42-51	~43-52	~42-51

C. 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 real-time, resource-efficient player recognition in sports environments.

D. Impact of Machine Learning Classifiers and Ensemble Methods

To address the remaining overfitting difficulties with ResNet50, a hybrid fusion technique was used, extracting features from ResNet50 and passing them to simpler machine learning classifiers under the hypothesis that the overfitting was caused by the classification layers in these deep learning architectures. Initial experiments with classifiers such as SVM, KNN, Random Forest, Gradient Boost Machine, and Decision Tree revealed that SVM and KNN performed well, enhancing classification accuracy while reducing overfitting. The following outcomes summarize the improvement:

- **ResNet50 + SVM**: Achieved an accuracy improvement of 95.78% with reduced overfitting compared to standalone ResNet50.
- **ResNet50 + KNN**: Provided competitive accuracy with stable recall and precision, achieving near SVM performance.

TABLE II
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

E. 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 III summarizes the ensemble approaches' outcomes.

TABLE III
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

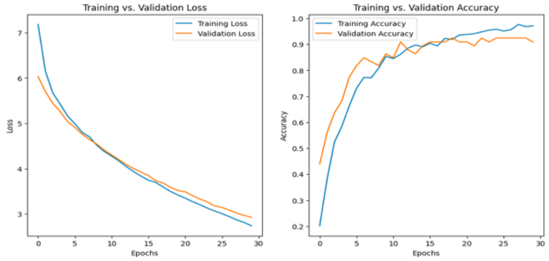
Performance Visualization

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. 4) 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.

Fig. 4. Training vs. Validation Loss and Accuracy for Baseline ResNet50 Model.



Stacking Ensemble: The learning curve for the stacking ensemble (Fig. 5) 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.

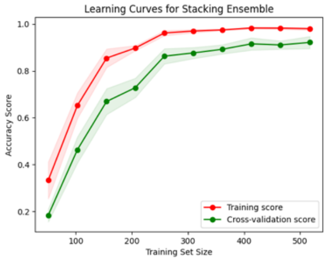


Fig. 5. Learning Curve for Stacking Ensemble Method.

2. Confusion Matrix Comparison:

Baseline Model (ResNet50): The confusion matrix for ResNet50 (Fig. 6) 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.

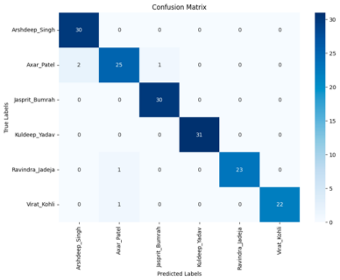


Fig. 6. Confusion Matrix for Baseline ResNet50 Model.

Stacking Ensemble: The confusion matrix for the stacking ensemble (Fig. 7) 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.

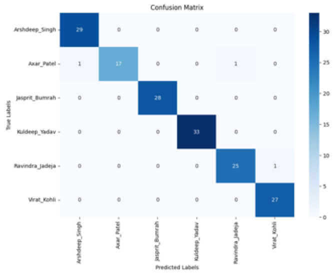


Fig. 7. Confusion Matrix for Stacking Ensemble Model.

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. The baseline ResNet50 model, as shown in Fig. 4, exhibited overfitting, with a significant difference in training and validation accuracy. However, the stacked ensemble learning curve (Fig. 5) showed significant generalization. Furthermore, the confusion matrix for the stacking ensemble (Fig. 7) shows better classification accuracy, with more True Positives, significantly fewer False Positives and False Negatives than the baseline ResNet50 model (Fig. 6). These findings validate the effectiveness of combining deep learning feature extraction with machine learning classifiers for enhanced spatial recognition.

V. 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.

A. 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.

Furthermore, the improved computational efficiency of the hybrid approach provides significant advantages for real-time applications. While effective in detailed feature extraction, Vision Transformers, proved to be computationally impractical for this project’s requirements due to its extended training times and high resource demands. In contrast, ResNet50 combined with machine learning classifiers offered a practical solution that balanced performance and efficiency, which is critical for sports analytics systems that require real-time or near-real-time processing.

Additionally, during dataset validation using the LLM evaluation method, minor deviations in the model's predictions for two players were observed when compared to human observations, during the error analysis. These deviations, most likely caused by subtle visual similarities between the players, such as overlapping features or similar stances, highlight the difficulties of recognizing visually similar individuals in dynamic match conditions. Despite this, the hybrid spatial fusion model's design effectively mitigates such discrepancies, relying on its robust stacking ensemble to maintain high accuracy and generalizability. This demonstrates the proposed approach's resilience and reliability in dealing with real-world complexities.

B. 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. However, its completeness and representation of various cricketing scenarios could benefit from expert evaluation such as feedback from professional cricket analysts or coaches which could help ensure that the dataset covers diverse playing conditions, player behaviours, and match scenarios, making the model more useful and reliable in real-world applications.

Lastly, while the hybrid model produced robust results, there are instances with considerable occlusion or overlapping players still presenting challenges. In such cases, to further improve recognition accuracy, addressing these issues may require integrating additional data modalities, such as temporal movement patterns, or exploring more advanced ensemble strategies.

VI. Conclusion

The proposed method presents a novel approach to player recognition in T20 cricket by combining deep learning feature extraction with machine learning classifiers in a hybrid spatial fusion recognition model. This model utilizes ResNet50 for feature extraction and integrates SVM, KNN classifiers, and a final Logistic Regression layer through a stacking ensemble method to improve robustness and accuracy under challenging real-world conditions. It is a hybrid approach that addresses the limitations of traditional deep learning models, specifically issues of overfitting and high computational costs, which are prevalent in complex environments with variable lighting, occlusions, and distant camera angles.

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 incorporating temporal data, such as movement patterns of players, 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. Addressing current dataset limitations through expert evaluation by professional cricket analysts and coaches will ensure diverse coverage of playing conditions and enhance model training. 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.