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 | 3rd Given Name Surname  *dept. name of organization (of Aff.) name of organization (of Aff.)*  City, Country email address or ORCID |

*Abstract*—This document is a model and instructions for LATEX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. \*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

*Index Terms*—component, formatting, style, styling, insert. *Abstract*—This document is a model and instructions for LATEX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. \*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

*IndexTerms*—component, formatting, style, styling, insert.*Abstract*—This document is a model and instructions for LATEX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. \*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract. *Abstract*—This document is a model and instructions for LATEX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. \*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

*Index Terms*—component, foratting, style, styling, insert.

# 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 Machines (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 Support Vector Machines (SVM) and K-Nearest Neighbors (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.

# II. RELATED WORK

*A. Traditional Deep Learning Approaches in Sports Analytics*

Player recognition and tracking in sports analytics have received increasing attention in recent years, particularly with the advancements in computer vision and deep learning. Traditional approaches have primarily relied on Convolutional Neural Networks (CNNs) for feature extraction and classification. CNN-based models are widely adopted for recognizing players in dynamic sports environments due to their ability to identify spatial patterns in images [1]. However, CNNs are often limited by their tendency to overfit on small datasets and struggle with complex, real-world scenarios such as sports fields with frequent occlusions and varying lighting conditions [2].

*B. Advanced Deep Learning Architectures*

Advanced architectures, such as ResNet and Vision Transformers, have further pushed the boundaries of deep learning in image recognition. ResNet’s residual learning capability has proven effective in mitigating issues related to gradient vanishing, making it suitable 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 less feasible for real-time applications [5], [6]. Despite their advantages, these models often overfit when applied to limited, domain-specific data, such as that found in sports analytics.

*C. Hybrid Approaches Combining Deep Learning and Machine Learning*

In response to the limitations of standalone deep learning models, recent research has explored hybrid approaches that integrate deep learning feature extractors with machine learning classifiers to improve model robustness and generalizability. Özyurt (2020) demonstrated that combining deep learning models with simpler machine learning classifiers, such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), can mitigate overfitting and improve classification performance in complex environments [7]. Similarly, Kibriya et al. (2021) applied a CNN-SVM hybrid model for brain tumor classification, showing that integrating a deep feature extractor with a machine learning classifier can enhance model accuracy and reduce computational demands in medical imaging tasks [8], [9].

*D. Contribution of This Work*

This paper builds on these findings by developing a spatial recognition model that fuses ResNet50 with machine learning classifiers. By applying a stacking ensemble of SVM and KNN classifiers and a final Logistic Regression layer, our approach addresses the overfitting issues commonly found in deep learning models and provides an efficient, real-time solution for recognizing cricket players in variable and challenging field conditions.

# III. METHODOLOGY

This section outlines the overall design of the proposed hybrid spatial fusion recognition system (Fig. 1). This framework outlines the key steps, from data acquisition and preprocessing to model training and player classification, providing a high-level overview of the system's structure and workflow. The objective was to create a robust player recognition model that effectively addresses overfitting and achieves reliable performance in dynamic cricket environments.

A diagram of a process flow

Description automatically generated

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

Player recognition in sports analytics, particularly in the dynamic environment of Twenty20 Internationals (T20i) cricket, presents unique challenges that hinder the accuracy and consistency of traditional recognition models. Real-time identification of fielders during the final overs of a T20 match is critical, as it directly influences both strategic decisions and statistical analyses. However, variable lighting conditions, frequent occlusions caused by overlapping players, and distant camera angles significantly complicate accurate recognition. Traditional computer vision and deep learning models, such as CNNs and Vision Transformers, have demonstrated capability in object and pattern recognition but fall short in highly variable environments. These models are particularly susceptible to overfitting on limited training data, reducing their ability to generalize under real-world conditions. Additionally, the computational requirements for high-performing models, like Vision Transformers, make them impractical for real-time applications.

To address these challenges, this research develops a spatial recognition model that combines **ResNet50 for feature extraction** with **machine learning classifiers** such as SVM, KNN and a final layer of Logistic Regression within a stacked ensemble model approach for classification. This hybrid approach leverages the feature extraction power of deep learning while enhancing robustness and reducing overfitting through simpler classifiers, resulting in improved accuracy and efficiency under challenging conditions.

*A. Data Pre-processing*

To accurately isolate players from the background, the YOLOv3 model was employed for player detection as shown in Fig. 1, the data preprocessing phase includes YOLOv3 for player detection and CLAHE for image enhancement, ensuring that only high-quality, focused player images are fed into the model.. YOLOv3’s ability to perform real-time object detection with high accuracy makes it well-suited for segmenting players from each frame in video footage. However, due to multiple boxes often detected in each frame, an additional selection criterion was applied: the largest vertical bounding box was selected, as the player’s upright posture while running or walking makes them the most prominent vertical object in the frame. And CLAHE (Contrast Limited Adaptive Histogram Equalisation) was used to improve picture quality before ResNet50 was used to extract features (Fig. 1). This approach minimized background noise and focused detection on the player.

*B. Model Selection and Evaluation*

The initial stage of model selection involved testing various deep learning architectures, starting with a custom CNN model. However, this approach led to poor performance, prompting the need for more sophisticated architectures. Several models were subsequently evaluated, including DenseNet, EfficientNet, Inception, MobileNet, VGG16, Xception, NASNet 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 tested, leveraging five different ViT models across multiple epochs and different configurations. Despite their potential, ViTs proved to be computationally expensive, requiring substantial training time, approximately 100 mins per epoch. Furthermore, they were prone to overfitting on the limited dataset available, making them unsuitable for real-time recognition in this context. Consequently, ResNet50 was selected as the primary feature extraction model due to its balance between performance and computational efficiency.

While ResNet50 showed improvements over previous models, overfitting remained despite early stopping and parameter adjustment. This led to the hypothesis that the overfitting was caused by the classification layers in these architectures. To solve this, a fusion strategy was taken: ResNet50 was utilized only for feature extraction, while SVM and KNN were employed for classification. The extracted features were classified using a stacking ensemble of SVM, KNN, with a final Logistic Regression classification layer. Hyperparameter tuning was done to all components, which resulted in effective overfitting mitigation and better model performance.

The detailed architecture of the hybrid spatial fusion recognition model is illustrated in Fig. 2. This diagram highlights the specific components and interactions, including the feature extraction layer (ResNet50), base classifiers (SVM and KNN), and the meta-classifier (Logistic Regression). Each step in the architecture is designed to mitigate overfitting while ensuring robustness in challenging conditions.

A diagram of a process

Description automatically generated

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

*C. Stacking Ensemble for Classification*

To enhance classification performance, several machine learning classifiers were evaluated, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Gradient Boost Machine, and Decision Tree classifiers. Among these, **SVM** and **KNN** delivered the most promising results, reducing overfitting and providing consistent classification accuracy. However, to further improve robustness, a stacking ensemble method was introduced.

The proposed method implemented a stacking ensemble architecture that combines the strengths of machine learning classifiers to complement the deep learning feature extractor. The stacking ensemble integrates Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) as base classifiers, with **Logistic** **Regression** serving as the meta-classifier. By leveraging the outputs of multiple classifiers, the stacking ensemble enhances robustness and generalizability, effectively mitigating overfitting and improving performance in challenging conditions.

Features were extracted from preprocessed player images using ResNet50, a deep learning architecture renowned for its residual learning capabilities. ResNet50 was used with its classification layer removed, ensuring the extracted feature vectors were optimized for classification. These features were then normalized and underwent dimensionality reduction to ensure compatibility with the subsequent machine learning classifiers. Fig. 2 elaborates on the flow of features from ResNet50 to the stacking ensemble, detailing how base classifiers (SVM and KNN) work in tandem with the meta-classifier (Logistic Regression) to produce a unified prediction.

**Workflow of the Stacking Ensemble**

**Base Classifiers:**

* **Support Vector Machine (SVM):** Operates with an RBF kernel, focusing on separating non-linear patterns in the feature space extracted by ResNet50. SVM mitigates noise and enhances the model's robustness to variations in data.
* **K-Nearest Neighbors (KNN):** Captures local spatial relationships in the feature space using the Manhattan distance metric. Its ability to generalize well complements the precision of SVM, creating a balanced and effective foundation for classification.

**Meta-Classifier:**

* **Logistic Regression:** Serves as the final decision-making layer, combining predictions from SVM and KNN. The meta-classifier refines these predictions into a unified output, reducing individual classifier errors and improving 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.

Various ensemble techniques, including Voting Classifiers and Decision-Level Fusion, were evaluated. However, the stacking ensemble method outperformed these approaches in accuracy and robustness. By combining the strengths of ResNet50 for feature extraction with SVM and KNN for base classification and Logistic Regression as the meta-classifier, the stacking ensemble achieved superior results under challenging conditions such as low-light environments, occlusions, and distant camera angles.

Tuned hyperparameters and cross-validation further minimized overfitting, resulting in a model that effectively addresses the unique challenges of T20 cricket player recognition. This stacking ensemble approach not only enhances classification accuracy but also provides a computationally efficient and scalable solution, demonstrating its potential for broader applications in sports analytics.

# IV. EXPERIMENTS

In this section, we present the results obtained from the spatial model’s evaluations, comparing various architectures and configurations to assess the performance and robustness of the final approach. To evaluate the performance of the spatial recognition models in the context of T20 cricket player recognition, the following metrics were used:

1. **Accuracy**: Accuracy measures the overall correctness of the model by comparing the number of correctly identified players to the total number of players evaluated. It is particularly useful for assessing general model performance in scenarios where both positive and negative predictions matter equally.

A black text with black text

Description automatically generated with medium confidence

1. **Precision**: Precision measures the exactness of the model by evaluating how many of the predicted players were actually correct. In player recognition, high precision ensures that false detections (e.g., background objects identified as players) are minimized.

A black and white math equation

Description automatically generated

1. **Recall**: Recall measures the completeness of the model by evaluating how many of the actual players present in the frames were correctly identified. This is critical in cricket analytics to ensure that no player goes unrecognized during high-stakes moments.

A mathematical equation with black text

Description automatically generated

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 tailored for cricket player recognition under dynamic match conditions, a novel dataset was developed specifically for the proposed method. The dataset comprises annotated images of six cricket players, each represented by 120–150 images captured from various angles and match scenarios. These images were meticulously extracted from high-resolution T20 match footage and manually labeled to ensure accurate player identification.

The dataset is designed to represent the following challenging conditions:

1. **Diverse Angles**: Images captured from multiple camera perspectives to simulate real-world field conditions and test the model’s robustness against angle variability.
2. **Lighting Variations**: Frames taken under varying lighting conditions, including daylight, dusk, and artificial lighting, to mimic the dynamic environment of T20 matches.
3. **Occlusions and Overlaps**: Instances where players are partially obscured by other players or objects, emphasizing the need for robust recognition systems.

To enhance the visual clarity and consistency of the dataset, preprocessing techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE) were applied. This step improved image quality by mitigating the effects of poor lighting and ensuring better feature extraction during model training.

**Validation of the Dataset**

The credibility of the novel cricket player recognition dataset was established through **LLM-based validation** using the pretrained **CLIP model**. The validation process involved evaluating the model's ability to distinguish between six cricket players using embeddings generated for each image in the dataset. This was performed under diverse conditions without modifying the dataset, ensuring its integrity.

**Results**

1. **Overall Accuracy:** The model achieved an accuracy of **82.00%** after applying a confidence threshold, highlighting the dataset's robustness in supporting recognition tasks under dynamic conditions. Predictions with confidence scores below the threshold were classified as "uncertain," reducing misclassifications.
2. **LLM Evaluation Graph:** A prediction distribution graph (Fig. 3) was generated to evaluate the model's output across all categories. The majority of uncertain predictions corresponded to the challenging classes (**Arshdeep Singh** and **Kuldeep Yadav**), supporting the effectiveness of the thresholding approach.

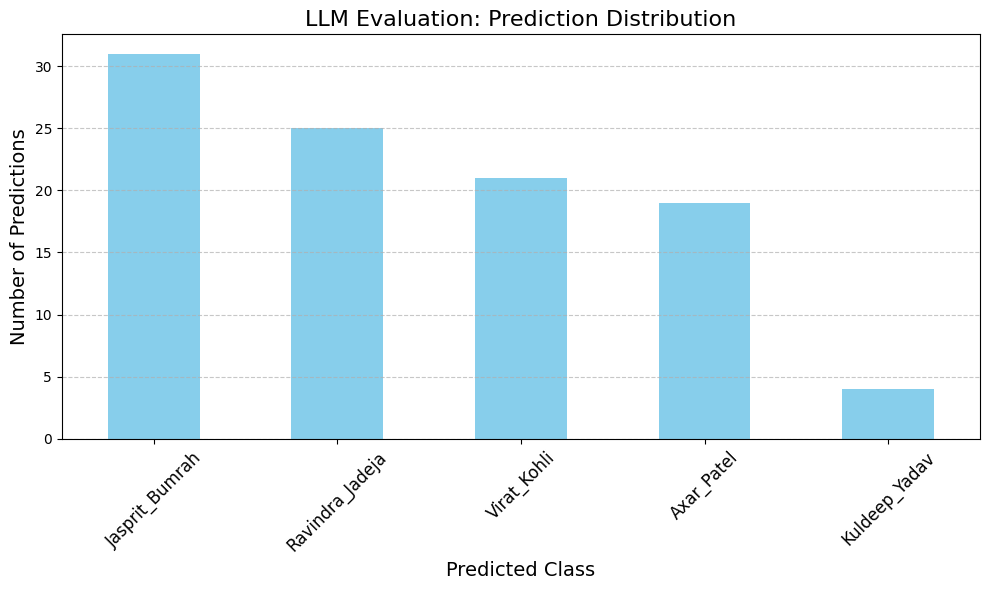


Fig. 3. LLM Evaluation grpah of the Prediction Distribution.

These results affirm the credibility of the dataset as a reliable resource for cricket player recognition tasks under real-world conditions. Its validation underscores its suitability for use in this project and further research in sports analytics.

*B. Model Comparison and Performance Metrics*

To identify the most effective model for player recognition, we initially compared several deep learning architectures, including a custom CNN, DenseNet, EfficientNet, Inception, MobileNet, VGG16, Xception, NASNet, and ResNet50. Each model was evaluated using accuracy, precision and recall, with ResNet50 emerging as the top performer across all metrics. The following summarizes the results of the initial model comparisons:

* **Custom CNN**: Showed poor accuracy and high overfitting, making it unsuitable for complex cricket environments.
* **DenseNet, EfficientNet, Inception, MobileNet, VGG16, NASNet, Xception**: These models performed moderately well but exhibited limited generalizability, particularly in low-light and occluded scenarios.
* **ResNet50**: Demonstrated superior feature extraction capabilities and robustness, achieving high accuracy with less 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 tested on the dataset, utilizing five different ViT models trained across multiple epochs. Despite the promise shown in other domains, ViTs proved computationally expensive, with a single epoch requiring approximately one hour and forty minutes. Furthermore, they showed a high tendency to overfit, achieving minimal improvements in accuracy over ResNet50 but at a significant computational cost. This confirmed 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 mitigate the remaining overfitting issues with ResNet50, a hybrid fusion approach was adopted, extracting features from ResNet50 and passing them to simpler machine learning classifiers based on the hypothesis that the overfitting was caused by the classification layers in these architectures. Initial trials with classifiers such as SVM, KNN, Random Forest, Gradient Boost Machine, and Decision Tree highlighted that **SVM** and **KNN** achieved the best results, improving classification accuracy while reducing overfitting. The following results 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 method, combining ResNet50 with SVM and KNN with a final Logistic Regression classification layer, demonstrated the highest robustness and accuracy among all configurations tested. By leveraging the strengths of multiple classifiers, the stacking ensemble effectively addressed overfitting while maintaining high generalizability across various conditions, including low-light and occluded player scenarios.

Various ensemble techniques were considered, including Voting Classifiers and Decision-Level Fusion, but the stacking ensemble method outperformed them, demonstrating superior accuracy and robustness. The results of the ensemble techniques are summarized in Table III:

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 further validate the performance of the stacking ensemble, the results were compared with 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.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generatedFig. 4. Training vs. Validation Loss and Accuracy for Baseline ResNet50 Model.

**Stacking Ensemble**: The learning curve for the stacking ensemble (Fig. 5) demonstrates minimal overfitting, with training and cross-validation accuracy scores converging as the training size increases. This highlights the ensemble’s ability to generalize effectively across varied conditions, including low-light and occluded environments.

A graph of a number of red and green lines

Description automatically generated

Fig. 5. Learning Curve for Stacking Ensemble Method.

1. Confusion Matrix Comparison:

**Baseline Model (ResNet50)**: The confusion matrix for ResNet50 (Fig. 6) reveals several misclassifications, particularly for players with similar jersey numbers or occluded by other players. Higher False Positives (FP) and False Negatives (FN) indicate that the baseline model struggled with complex scenarios.

A graph with blue squares and white text

Description automatically generated

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 a direct result of the ensemble’s robust classification capabilities.

A graph with blue squares and white text

Description automatically generated

Fig. 7. Confusion Matrix for Stacking Ensemble Model.

These results indicate that the stacking ensemble method, with tuned hyperparameters and cross-validation, provided the best performance, achieving a significant reduction in overfitting and delivering reliable player recognition in real-world conditions. The baseline ResNet50 model, as shown in Fig. 4, exhibited overfitting, with noticeable divergence between training and validation accuracy. However, the stacking ensemble learning curve (Fig. 5) demonstrated strong generalization. Furthermore, the confusion matrix for the stacking ensemble (Fig. 7) highlights improved classification accuracy, with higher True Positives and significantly fewer False Positives and False Negatives compared to 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

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

*A. Benefits of the Hybrid Approach*

The hybrid model with ResNet50 and machine learning classifiers achieved notable improvements in performance metrics, particularly in accuracy and robustness. This approach addresses one of the main limitations of deep learning models—overfitting on small or domain-specific datasets—by utilizing simpler classifiers that generalize well without requiring 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 might produce, especially in scenarios with limited training data.

Moreover, the computational efficiency of the hybrid approach is a significant advantage for real-time applications. Vision Transformers, while effective in detailed feature extraction, proved computationally impractical for this project’s requirements due to extended training times and high resource demands. In contrast, ResNet50 combined with machine learning classifiers offered a practical solution, balancing performance with efficiency, which is essential for sports analytics systems requiring real-time or near-real-time processing.

*B. Limitations and Trade-offs*

Despite the advancements achieved, some limitations remain. The stacking ensemble approach, while reducing overfitting, can introduce complexity in model deployment due to the integration of multiple classifiers. Additionally, although cross-validation and hyperparameter tuning improved model robustness, the ensemble method may still require further tuning to adapt to different sports or environmental conditions, such as varying field sizes or camera angles specific to each venue.

Another limitation involves the dataset itself. The dataset was designed and annotated for this study, but its completeness and representation of various cricketing scenarios could benefit from expert evaluation. Feedback from professional cricket analysts or coaches could help ensure the dataset covers diverse playing conditions, player behaviors, and match scenarios, making the model more practical and reliable in real-world applications.

Lastly, while the hybrid model demonstrated robust results, instances with significant occlusion or overlapping players still present challenges. Addressing these issues may require integrating additional data modalities, such as temporal movement patterns, or exploring more advanced ensemble strategies to further improve recognition accuracy in such cases.

# 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. The model leverages ResNet50 for feature extraction and integrates Support Vector Machine (SVM), K-Nearest Neighbors (KNN) classifiers, and a final Logistic Regression layer through a stacking ensemble method to improve robustness and accuracy under challenging real-world conditions. This hybrid approach addresses the limitations of traditional deep learning models, particularly issues of overfitting and high computational costs, which are prevalent in environments with variable lighting, occlusions, and distant camera angles.

Experimental results demonstrate that the stacking ensemble method effectively enhances model generalizability and performance, 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. The findings indicate that the proposed model not only meets the needs of T20 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, to complement the spatial recognition model and 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 research 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. Srinadh Bhojanapalli et al., "Understanding Robustness of Transformers for Image Classification," *arXiv*, 2021
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.