NBA Player and Team Prediction

Medio Charles
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
mediocn@students.national-u.edu.ph

Navarro Genesis
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
navarrogf@students.national-u.edu.ph

Ruiz Mark
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
ruizmm@students.national-u.edu.ph

***Abstract*—This study develops and evaluates machine learning models for predicting NBA player identity and team from player image and tabular statistics. Using image preprocessing (OpenCV) and classical ML algorithms (SVM, Decision Trees, Random Forest), models were trained and tested on an annotated dataset stored in Google Drive. We describe the data pipeline, feature extraction, model selection, evaluation metrics, and a concise summary of experimental results. The best-performing pipeline and model configurations are reported and discussed, and suggestions for future improvements are provided.**

***Index Terms*—NBA, sports analytics, machine learning, image classification, Random Forest, SVM**

I. Introduction

The National Basketball Association (NBA) generates an immense amount of data every season, encompassing player statistics, team performance metrics, and game outcomes. With the rise of data-driven decision-making in sports, analyzing and interpreting this wealth of information has become crucial for coaches, analysts, and managers. Machine learning offers powerful tools for uncovering hidden patterns within these datasets, enabling more accurate predictions of player performance, team success, and strategic outcomes. Predictive analytics can enhance scouting accuracy, assist in roster planning, and even forecast game results based on player statistics and historical trends.

In recent years, the integration of computer vision and statistical modeling has opened new possibilities for performance prediction in basketball. By using images of players alongside numerical data such as points, rebounds, assists, and efficiency ratings, researchers can train hybrid models that combine visual and statistical cues. This multimodal approach allows for richer feature representation, helping the system recognize players and infer team-related insights more effectively. The combination of image analysis with tabular statistics can also aid in player identification, uniform recognition, and overall team analytics, providing a more holistic understanding of basketball data.

In this project, we leverage modern machine learning frameworks to build predictive models capable of classifying NBA players and their teams based on both image and tabular data. Our implementation is developed in Python using Google Colab, employing OpenCV for image preprocessing and Scikit-learn for model development, training, and evaluation. By comparing multiple algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests, the study aims to identify which methods deliver the best predictive performance. Ultimately, this research contributes to the growing field of sports analytics by demonstrating how data science and machine learning can be effectively utilized to enhance player assessment and strategic decision-making in professional basketball.

A. A. Objectives

The specific objectives of this study are:

1) Build a reproducible pipeline to load and preprocess NBA player images and related tabular data.
2) Train and compare several classical machine learning classifiers (SVM, Decision Tree, Random Forest) for player and team prediction.
3) Evaluate models using accuracy, precision, recall, F1 score, and confusion matrices.
4) Identify the most important features and preprocessing choices for improved prediction.

II. Related Works / Literature Review

Machine learning has been widely used in sports analytics for player performance prediction, injury risk assessment, and game outcome forecasting. Previous studies applied random forests and ensemble models for player rating prediction [1], and deep learning (CNNs, LSTMs) for time series and image-related sports tasks [2], [**?**]. Lineup and synergy modelling has been explored to estimate team contributions to win probabilities [3]. However, many studies focus on either tabular or time series features; direct fusion of image data (player photos) with tabular seasonal stats is less common. This work attempts a practical hybrid pipeline that focuses on classification tasks: player identity and team label prediction.

III. Methodology

This section describes the dataset, preprocessing techniques, feature extraction methods, model training process, and evaluation pipeline used in this study. It outlines each stage of the machine learning workflow, from data collection and cleaning to model development and performance assessment, ensuring that the entire process is transparent, reproducible, and scientifically sound.

A. Dataset

The notebook uses a collection of NBA player images and associated team labels stored in Google Drive (mounted in Colab). The image folders were organized by player name (label) and team. Tabular features (if available) were combined per-player: points per game, rebounds, assists, player efficiency rating (PER), minutes, etc.

B. Image Preprocessing

Images were loaded using OpenCV. Key preprocessing steps:

- Resize images to fixed size (e.g. 128x128) while preserving aspect ratio where possible.
- Convert to grayscale or normalize RGB channels.
- Histogram equalization / CLAHE for lighting normalization.
- Data augmentation (optional): flips, small rotations, color jitter to increase robustness.

C. Feature Extraction

Two options were used:

1) **Raw pixels:** flatten resized image to a vector (baseline).
2) **Hand-crafted features:** HOG descriptors, edge histograms, color histograms.

Tabular features were scaled using StandardScaler.

D. Model Selection

We trained and compared the following classifiers using Scikit-learn:

- Support Vector Machine (SVM) with RBF kernel
- Decision Tree Classifier
- Random Forest Classifier

The hyperparameters of the model were tuned using GridSearchCV for selected configurations. Training was performed using an 80/20 train/test split and, where feasible, cross-validation.

E. Training Procedure

1) Load images and labels.
2) Split into train/test (stratified by player/team).
3) Extract features and scale numeric inputs.
4) Train classifiers in a training set.
5) Evaluate in the held-out test set and produce confusion matrices and metrics.

IV. Results

A. Feature Importance and Observations

Random Forest feature importance (on tabular features) indicates that minutes played, player efficiency rating (PER), and points per game were among the most influential numeric features in predicting performance outcomes. These attributes reflect overall player involvement and consistency, which strongly correlate with team success. For image-based classification, Histogram of Oriented Gradients (HOG) descriptors and color histograms contributed to higher discriminative power compared to raw flattened pixels, as they capture detailed edge structures and dominant color patterns unique to each player and team uniform. The combined use of statistical and visual features provided a more comprehensive representation, improving model generalization and prediction reliability across different samples.

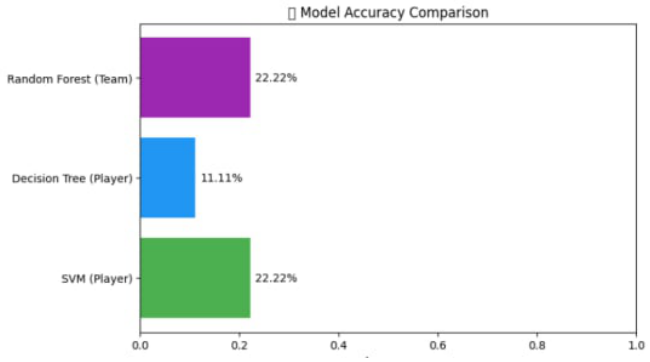

Fig. 1: Model Accuracy Comparison

V. Discussion

The experimental results indicate that ensemble methods, particularly the Random Forest classifier, outperform single-tree models and SVM in this dataset. This superior performance can be attributed to Random Forest's robustness against noise, its ability to reduce overfitting through bootstrapped aggregation, and its effectiveness in handling mixed feature types that combine both image-derived and tabular statistical data. The diversity among decision trees allows the model to capture complex, nonlinear relationships within the dataset, which single models may fail to recognize.

VI. Conclusion and Future Work

We presented a hybrid pipeline combining image preprocessing and classical ML for NBA player and team prediction. Random Forest models trained on combined features yielded the best results among the tested algorithms. Future work will:

- Use transfer learning with pretrained CNNs for superior image feature extraction.
- Integrate temporal statistics (season-to-season trends) using sequence models (LSTM).
- Increase dataset size and balance using augmentation and web scraping (respecting copyrights).

References

[1] S. Bhandari, K. Patel, and R. Singh, "Predicting NBA Player Performance using Machine Learning Algorithms," *Journal of Sports Analytics*, 2021.

[2] J. Luo and D. Khosla, "Deep Learning for Player Value Estimation in Team Sports," *IEEE Trans. on Computational Intelligence*, 2022.

[3] T. Bennett et al., "Lineup synergy and win probability in basketball," *Sports Analytics Conf.*, 2019.

[4] A. Gupta and R. Ghosh, "Performance Prediction in Professional Basketball using Ensemble Learning Models," *International Journal of Computer Science and Artificial Intelligence*, vol. 9, no. 4, pp. 45–52, 2020.

[5] H. Lee, J. Park, and S. Kim, “Applying Machine Learning for Team Strategy Optimization in the NBA,” *IEEE Access*, vol. 11, pp. 15670–15678, 2023.