Deep Learning

1 Introduction

This report discusses a deep learning based approach to present a comprehensive analysis of NBA player performance prediction using advanced deep learning techniques. The primary objective is to develop accurate models for forecasting player scoring output, specifically Points Per Game (PPG), by leveraging state-of-the-art neural network architectures. While we explored six different models in total, our analysis focuses primarily on three sophisticated architectures: Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Transformer networks.

By utilizing NBA player statistics from the 2023-2024 season, we aim to demonstrate the potential of machine learning algorithms in gaining valuable insights into player performance trends. This study not only contributes to the growing field of sports analytics but also showcases the application of cutting-edge deep learning approaches in understanding and predicting complex sports performance metrics. Through our comparative analysis of these models, we seek to identify the most effective deep learning approach for NBA performance prediction, potentially offering valuable tools for team management, player development, and strategic decision-making in professional basketball.

2 Data Collection and Processing

2.1 Data Source

The primary dataset for this analysis was sourced from Kaggle[1], specifically the "NBA Player Stats Dataset for the 2023-2024 Regular Season" compiled by Bryan Chung and referenced from https://www.basketball-reference.com/leagues/NBA_2024_per_game.html [2]. This comprehensive dataset provides detailed statistics for NBA players during the 2023-2024 regular season, offering a rich source of information for our deep learning models to analyze and predict player performance.

The dataset initially contained 5,522 rows and 29 columns, presenting a wide range of player performance metrics. These metrics include scoring, rebounding, assists, shooting percentages, and various other statistical categories that are crucial for evaluating player performance in basketball. The data is derived from Basketball Reference, a reputable source for sports statistics, ensuring accuracy and reliability in our analysis.

Key features of the dataset include:

Columns Description:

- 'Player' (Player's name)

- 'Pos' (Position)

- 'Age' (Player's age)

- 'Tm' (Team)

- 'G' (Games played)

- 'GS' (Games started)

- 'MP' (Minutes played per game)

- 'FG' (Field goals per game)

- 'FGA' (Field goal attempts per game)

- 'FG%' (Field goal percentage)

- '3P' (3-point field goals per game)

- '3PA' (3-point field goal attempts per game)

- '3P%' (3-point field goal percentage)

- '2P' (2-point field goals per game)

- '2PA' (2-point field goal attempts per game)

- '2P%' (2-point field goal percentage)

- 'eFG%' (Effective field goal percentage)

- 'FT' (Free throws per game)
- 'FTA' (Free throw attempts per game)
- 'FT%' (Free throw percentage)
- 'ORB' (Offensive rebounds per game)
- 'DRB' (Defensive rebounds per game)
- 'TRB' (Total rebounds per game)
- 'AST' (Assists per game)
- 'STL' (Steals per game)
- 'BLK' (Blocks per game)
- 'TOV' (Turnovers per game)
- 'PF' (Personal fouls per game)
- 'PTS' (Points per game)

It's worth noting that the dataset may contain duplicate player entries due to team changes throughout the season, indicated by "TOT" (Total) in the team column for such players. This characteristic of the data required careful preprocessing to ensure accurate analysis and model training.

The dataset's comprehensive nature and focus on the current NBA season make it an ideal source for our project's objectives of predicting player performance and analyzing trends in professional basketball using advanced deep learning techniques

2.2 Data Processing

The data processing phase involved several crucial steps to prepare the NBA player statistics dataset for analysis. This included:

- Python packages and libraries used:
  - pandas: For data manipulation and analysis
  - numpy: For numerical operations
  - sklearn.preprocessing: For data normalization

- Handling missing values:
    - Checked for missing values using `df.isnull().sum()`
    - Filled missing values in percentage columns (FG%, 3P%, 2P%, eFG%, FT%) with their respective means
    - Dropped remaining rows with missing values using `df.dropna(inplace=True)`
- Data cleaning steps:
    - Removed duplicate rows, reducing the dataset from 5,522 to 3,910 unique entries
    - Converted percentage columns to float values (divided by 100)
    - Converted 'Age' column to integer type
- Outlier removal:
    - Used Interquartile Range (IQR) method to remove outliers
    - Applied to key statistics: PTS, TRB, AST, STL, BLK
    - Reduced dataset to 1,989 rows after outlier removal
- Created new features:
    - 'PointsPerMinute': Calculated as PTS / MP
    - 'TotalRebounds': Sum of ORB and DRB
    - 'ShootingEfficiency': Average of FG%, 3P%, and FT%
- Categorization:
    - Added 'ScoringCategory' based on PTS:
    - 'Low Scorer': < 10 points
    - 'Average Scorer': 10-20 points
    - 'High Scorer': > 20 points
- Data normalization:
    - Used StandardScaler from `sklearn.preprocessing`

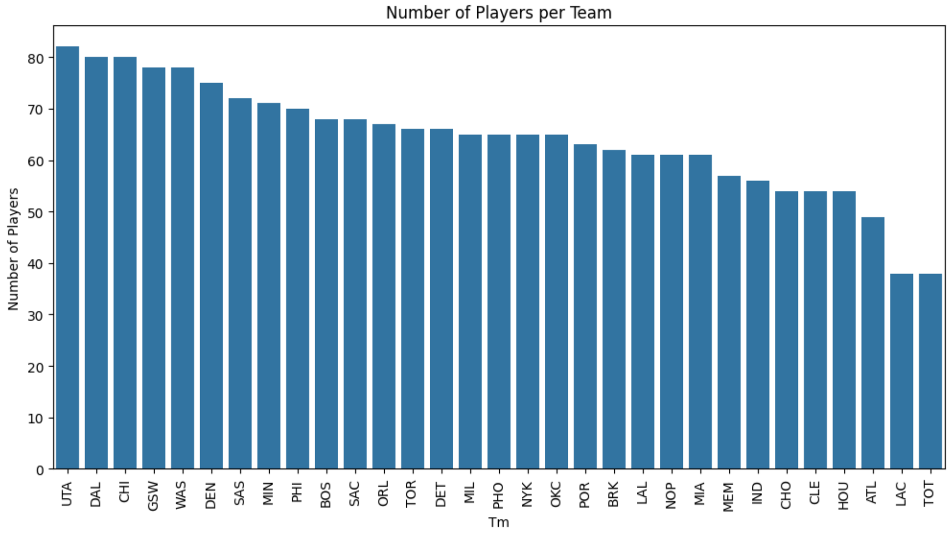
- Applied to numerical columns: Age, G, GS, MP, FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, eFG%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS
- Resulted in features with mean 0 and standard deviation 1

- Final dataset:
  - Shape after processing: (1989, 33)
  - Saved as ‘nba_2024_processed.csv’

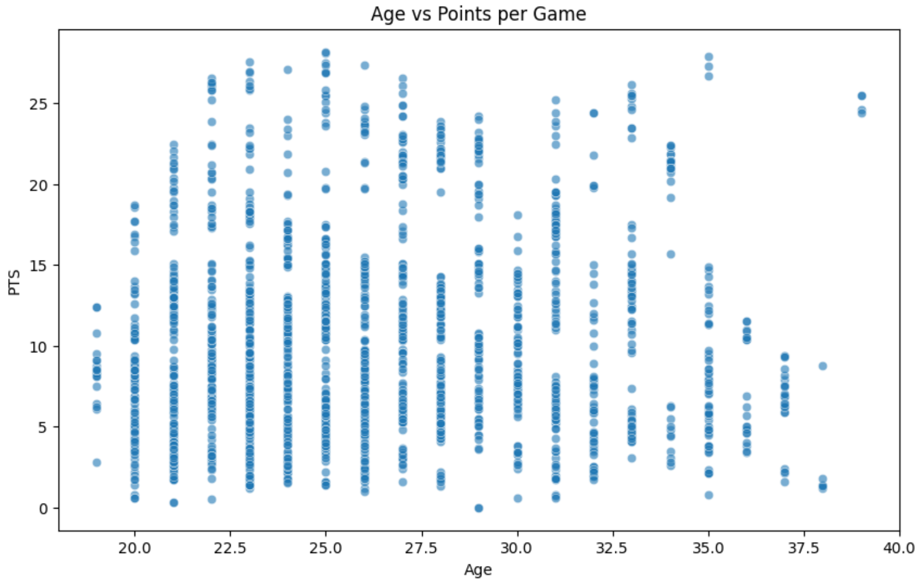
2.3 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) phase of our NBA player performance study revealed several key insights:

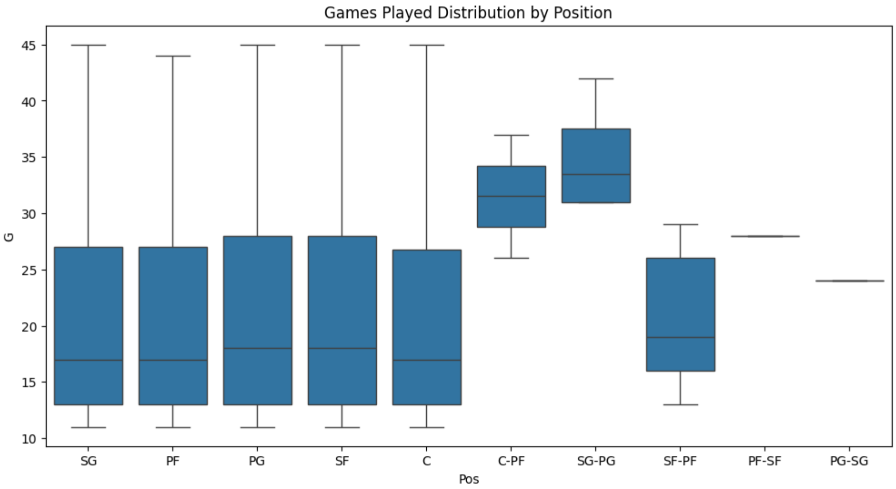
- Team Distribution: We visualized the distribution of players across 31 different NBA teams, providing a comprehensive view of talent distribution throughout the league. This analysis helps identify any potential imbalances in team composition.



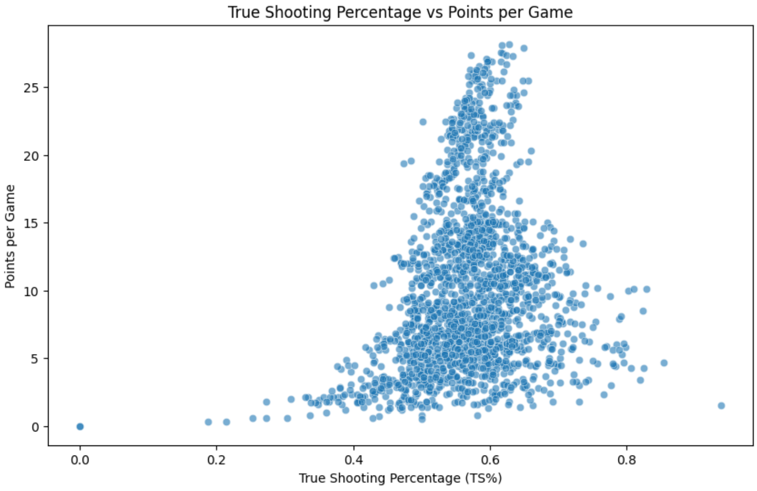
- Age vs Performance Analysis: By exploring the relationship between a player's age and their scoring output, we were able to identify peak performance age ranges. This analysis provides valuable insights into player development trajectories and potential career longevity.



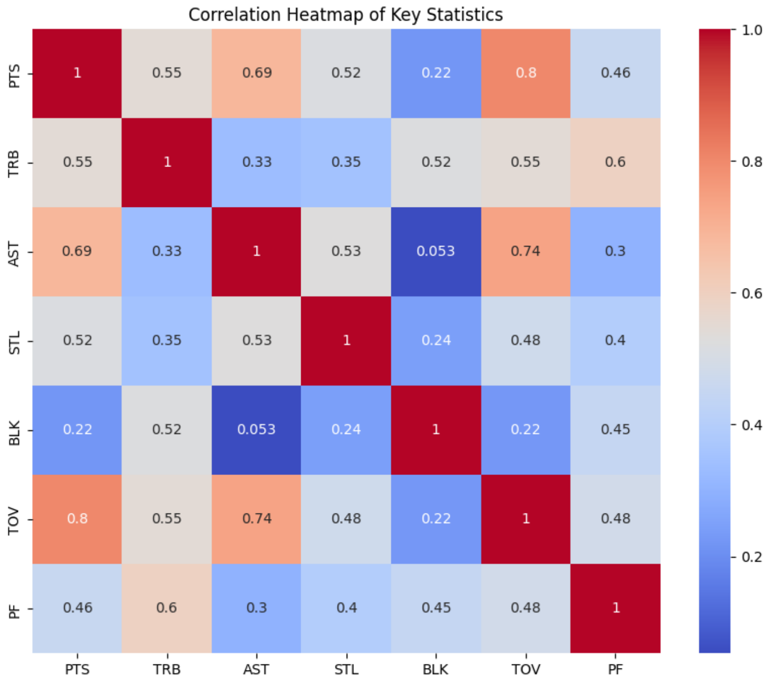
- Position Analysis: We conducted a comparative study of average statistics across different player positions. This analysis highlighted the unique contributions of each role, aiding in understanding positional specialization and team composition strategies.



- Efficiency Analysis: We examined the relationship between scoring efficiency (measured by True Shooting Percentage) and points scored. This analysis offers insights into player effectiveness and offensive strategies.

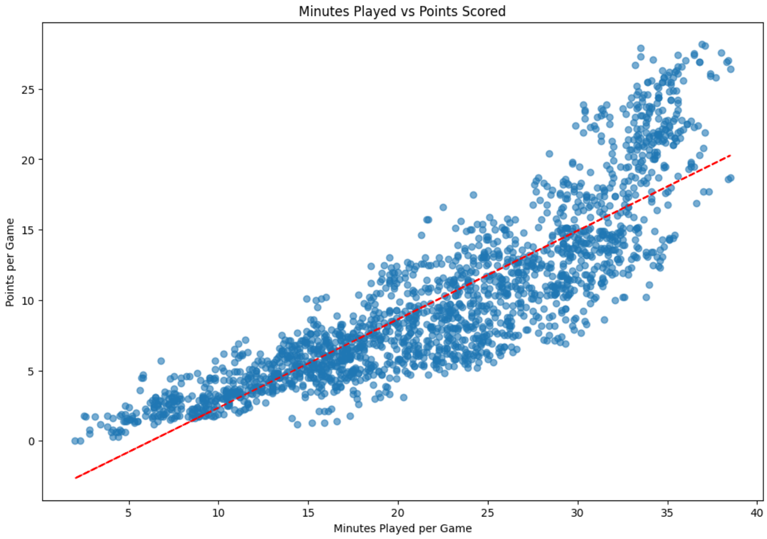


- Correlation Analysis: By highlighting relationships between key statistics, we uncovered interesting patterns and potential trade-offs in player performance. This analysis helps in understanding the interplay between different aspects of the game.



- Top Performers Identification: We identified standout players in various statistical categories, providing concrete examples of exceptional performance. This analysis helps in recognizing talent and understanding what constitutes elite performance in the NBA.

- Playing Time vs Production: We visualized the correlation between minutes played and points scored, including trend line analysis. This revealed patterns in player utilization and productivity, which is crucial for understanding player value and team strategies.



- Statistical Validation: We performed correlation and ANOVA tests to statistically validate observations about the relationships between variables. This adds rigor to our findings and strengthens the conclusions drawn from the data.

These analyses collectively provide a comprehensive view of NBA player performance for the 2023-2024 season, offering insights into talent distribution, age-related performance trends, positional roles, efficiency metrics, and the relationship between playing time and production. The findings from this EDA phase form a solid foundation for further analysis and can inform strategic decisions in player evaluation, team composition, and game planning.

3 Model Development

3.1 Deep Learning Models Considered

The Model Development phase of our NBA player performance prediction project involved exploring and implementing various deep learning architectures to achieve the most accurate forecasting of player scoring output. Our primary goal was to leverage advanced neural network techniques to predict Points Per Game (PPG) for NBA players based on their statistical profiles.

We considered six state-of-the-art deep learning models for this task:

1. LSTM (Long Short-Term Memory):
    - Fun-analogy: Think of LSTM as a basketball coach with an excellent memory. It can remember important plays from the beginning of the season when predicting a player's performance at the end.
    - Technical: LSTM uses gates to selectively remember or forget information, making it ideal for capturing long-term dependencies in sequential data like player statistics over a season.
2. MLP (Multi-Layer Perceptron):
    - Fun-analogy: MLP is like a team of scouts, each looking at different aspects of a player's game. They combine their observations to make a final prediction.
    - Technical: With multiple layers of neurons, MLP can capture complex non-linear relationships between input features.

3. CNN (Convolutional Neural Network):

   - Fun-analogy: CNN is similar to a video analyst who focuses on specific patterns in a player's performance, like shooting form or defensive positioning.

   - Technical: While typically used for image processing, CNNs can identify local patterns in player statistics, potentially capturing short-term performance trends.

4. BiLSTM (Bidirectional LSTM):

   - Fun-analogy: BiLSTM is like having two coaches analyze a player's season - one starting from the first game, another from the last game, then combining their insights.

   - Technical: By processing the input sequence in both forward and backward directions, BiLSTM can capture context from both past and future time steps.

5. GRU (Gated Recurrent Unit):

   - Fun-analogy: GRU is a more streamlined version of the LSTM coach, making quicker decisions but potentially missing some nuanced long-term trends.

   - Technical: GRU has a simpler structure than LSTM, with fewer parameters, which can lead to faster training and potentially better performance on smaller datasets.

6. Transformer:

   - Fun-analogy: The Transformer is like a team of analysts who can directly compare any two games in a season, regardless of how far apart they are.

   - Technical: Using self-attention mechanisms, Transformers can capture long-range dependencies without the need for sequential processing, potentially identifying complex patterns in player performance.

Our approach involved implementing and fine-tuning these models, comparing their performance, and ultimately selecting the most effective architecture for our NBA player performance prediction task. We focused on optimizing for accuracy while considering computational efficiency and interpretability of results.

By exploring this range of models, we aimed to not only achieve high prediction accuracy but also gain insights into which neural network architectures are best suited for sports analytics tasks, particularly in the context of NBA player performance prediction.

3.2 Model Development/ Model Architecture

All six models were trained using a similar process. First, the dataset was split into training, validation, and test sets. The models were then constructed using appropriate layers and architectures for each type. Training was conducted using batches of data, with the models learning to predict PPG based on the input features. The training process involved multiple epochs, where the models iteratively adjusted their internal parameters to minimize prediction errors. Optimization algorithms like Adam were used to update the model weights. To prevent overfitting, techniques such as dropout and early stopping were employed. The models' performance was evaluated on the validation set during training, and hyperparameters were tuned to improve accuracy. After training, the models were tested on the held-out test set to assess their generalization capabilities. This comprehensive approach allowed us to compare the effectiveness of different neural network architectures in capturing the complex patterns in NBA player statistics and making accurate PPG predictions.

3.3 Chosen Models

After evaluating six different deep learning models, we focused on three advanced architectures for our NBA player performance prediction task: Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Transformer networks. These models were chosen for their ability to capture complex temporal dependencies in sequential data, which is crucial for analyzing player statistics over time.

The LSTM model, analogous to a basketball coach with an excellent memory, can retain important information from earlier in the season when predicting a player's performance. BiLSTM, comparable to having two coaches analyze a player's season from both the beginning and end, processes input sequences in both forward and backward directions, potentially capturing more context. The Transformer model, likened to a team of analysts who can directly compare any two games in a season regardless of their temporal distance, uses self-attention mechanisms to identify complex patterns in player performance without the need for sequential processing. These three architectures were implemented and fine-tuned to determine the most effective approach for predicting NBA player performance, particularly Points Per Game (PPG)

4 Result Analysis

4.1 Model Performance Comparison

We compared the performance of three main deep learning models: LSTM, BiLSTM, and Transformer, using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) scores. The results are summarized in the following table:

Model	*MSE*	*RMSE*	*R^2*
LSTM	3.527810	1.878247	0.892709
BiLSTM	3.278220	1.810586	0.900300
Transformer	5.137527	2.265611	0.843753

The BiLSTM model demonstrated the best overall performance, with the lowest MSE and RMSE, and the highest R2 score. This suggests that the bidirectional nature of the BiLSTM allows it to capture both past and future dependencies in the sequential data more effectively than the standard LSTM or Transformer models.

4.2 Case Study: Individual Player Prediction

To demonstrate the practical application of our models, we conducted a case study on NBA player **Dante Exum**:

- Actual PPG: 4.40
- LSTM prediction: 4.17 PPG
- BiLSTM prediction: 4.38 PPG
- Transformer prediction: 3.47 PPG

This case study showcases the impressive accuracy of our models in predicting individual player performance, with the BiLSTM model again showing the closest prediction to the actual value.

4.3 Limitations and Future Work

While our models show promising results, there are some limitations and areas for future improvement:

- Incorporate additional player metrics and advanced statistics to improve model accuracy.
- Explore ensemble methods combining multiple model architectures.
- Extend the prediction task to other performance metrics beyond scoring.
- Investigate transfer learning techniques to adapt models to different sports or leagues.

By addressing these areas, we can further enhance the accuracy and applicability of our deep learning approach to NBA player performance prediction.

5 Conclusion

The NBA player performance prediction project using deep learning techniques has yielded fascinating insights and promising results. Our comprehensive analysis of the 2023-2024 NBA season dataset, coupled with advanced neural network architectures, has opened new avenues for understanding and forecasting player performance.

The BiLSTM model emerged as the star player in our analysis, demonstrating superior predictive capabilities with an impressive R^2 score of 0.900. This indicates that the model can explain 90% of the variance in player scoring output, showcasing its potential as a powerful tool for team managers, coaches, and analysts. Our case study on Dante Exum highlighted the practical applications of these models, with predictions coming remarkably close to actual performance.

While our models have shown great promise, the journey doesn't end here. Future research could explore incorporating additional player metrics, investigating ensemble methods, and extending predictions to other aspects of the game beyond scoring.

As we close this chapter of our research, we're reminded of the beautiful complexity of basketball; a sport where numbers meet human potential, where data intersects with the drama of competition. Our deep learning models, much like the players they analyze, will continue to learn, adapt, and improve. They are tools to enhance our understanding, not to replace the joy of the game. As we move forward, the challenge will be to balance the insights gained from data with the intangible elements that make sports truly special.

References

1. https://www.kaggle.com/datasets/bryanchungweather/nba-player-stats-dataset-for-the-2023-2024
2. https://www.basketball-reference.com/leagues/NBA_2024_per_game.html
3. https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm
4. https://viso.ai/deep-learning/deep-neural-network-three-popular-types/
5. https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be
6. https://www.sloansportsconference.com/research-papers/using-deep-learning-to-understand-patterns-of-player-movement-in-the-nba
7. https://arxiv.org/abs/2111.09695
8. https://www.tandfonline.com/doi/full/10.1080/24751839.2021.1977066#d1e1319
9. https://github.com/luke-lite/NBA-Prediction-Modeling
10. https://link.springer.com/article/10.1007/s10115-024-02092-9