A Deep Learning Approach to Predicting NBA MVP Winners

Team Name: Trophy Trackers

Team Members: Naman Tellakula, Vedanth Sathwik Toduru Madabushi

February 14, 2025

1 Project Summary

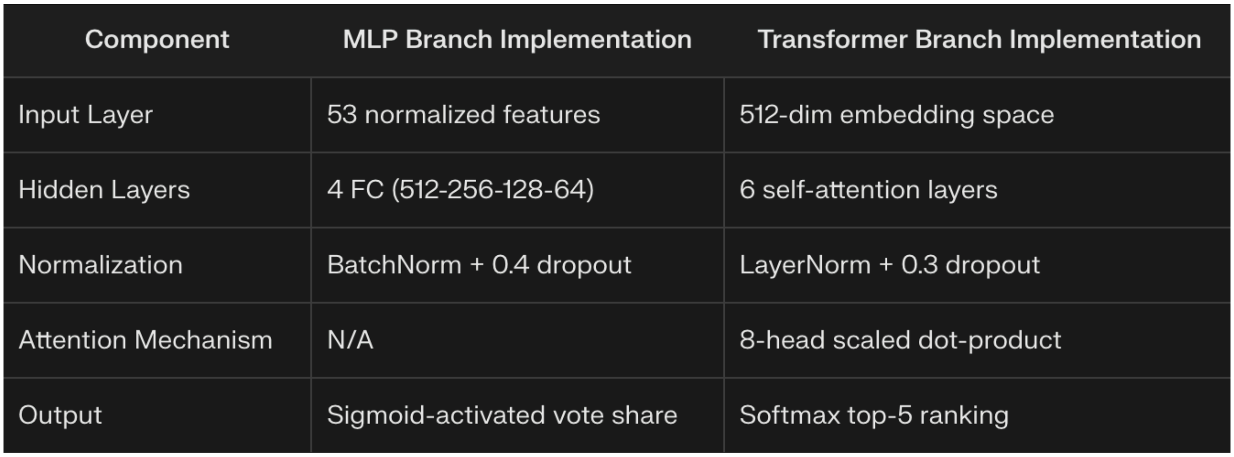
The NBA MVP award is one of the most prestigious awards in professional basketball, given to the best-performing player in the league each season. Determining the NBA MVP each year is an intricate yet thorough process in which voters consider various factors like impact, statistics, team performance, healthiness, and others to anoint one player as the best in basketball. The NBA MVP award symbolizes greatness and leadership in a league that has historically marketed its extraordinary individuals, so the “correct” person must be selected for this prestigious award. While the award's winner is voted on by a panel of 100 sportswriters and broadcasters, in this project, we seek to determine if there is a more objective way to identify the winner so that there is no bias, public perception, or other external factors that play a part. We plan to train different neural networks on all player and team data encompassing many statistics to predict the NBA MVP for each year and contribute meaningfully to the growing field of sports analytics.

This project aims to develop an objective deep-learning model to predict MVP outcomes, minimizing subjective biases inherent in human voting.

2 Approach

We plan to use PyTorch's multi-layer perceptron and transformer-based models for our deep learning analysis and predictions. Our dataset consists of various statistics, and these models are well-equipped to analyze such information and the relationships between different columns. Essentially, the input layer will receive all the relevant player statistics, such as points, rebounds, assists, etc., advanced metrics like player efficiency rating and win shares, and team data, like seeding and overall record. We will implement fully connected layers to capture the non-linear relationships between features for the hidden layers. Then, the output layer will be a single neuron that will predict the MVP voting share.

A possible implementation of the multi-layer perceptron and transformer-based model would look like this:



Component	MLP Branch Implementation	Transformer Branch Implementation
Input Layer	53 normalized features	512-dim embedding space
Hidden Layers	4 FC (512-256-128-64)	6 self-attention layers
Normalization	BatchNorm + 0.4 dropout	LayerNorm + 0.3 dropout
Attention Mechanism	N/A	8-head scaled dot-product
Output	Sigmoid-activated vote share	Softmax top-5 ranking

To evaluate the effectiveness of these models, we will use mean-squared error or mean-absolute error for the loss function and assess how many MVPs the model correctly identified. Our dataset contains player data from 1982 to 2022, so we will divide the number of correctly identified players by 41 to evaluate the model's accuracy.

We plan to experiment with the batch size, regularization values, optimizer, neurons, learning rate, etc., and hyper-tune these parameters to achieve peak model performance. We will analyze the graph of the loss and accuracy functions to ensure the model doesn't overfit or underfit, and we will continually tweak our hyperparameters until everything is aligned. Another way we can ensure the model doesn't overfit is by implementing early stopping.

The data visualizations will indicate whether the model is strong or not. However, some potential issues that could arise with the models include the lack of sufficient data to train on since we only have 41 different MVPs. If this becomes a problem, we can either look for a more comprehensive dataset or alter our project to predict the MVP shares of the top five finishers for the award each year, resulting in more model analysis.

Another issue might be that different eras (or decades for this purpose) emphasize specific statistics differently. If this is the case, we will split up the dataset and train different models for each of these eras. These eras would include the 1980s (Physical Play Era), 1995-2010 (Isolation Era), and 2011-present (Analytics/3PT Era) [4].

3 Related Work

There is some work that has been done in this area that we plan to build off. Researchers [1] have trained machine learning models like ElasticNet, Random Forest, and Gradient Boosting on player data but have achieved relatively low and inconsistent accuracies. Other studies [2] have relied on linear regression and decision trees with traditional stats, not incorporating other complex models, advanced metrics, or team data. There is little research into how MLP or transformer-based models perform on such data, especially with the extent of our dataset and statistics. We seek to expand on this research and determine if such models perform better than the traditional machine learning models.

4 Datasets

From Kaggle, our primary dataset [3] encompasses detailed player statistics from 1982-2022. This includes traditional statistics, advanced metrics, and win percentages. This dataset also contains each player's MVP share, which will be used to test the model. One issue with the dataset is its comprehensiveness because it includes data for all players. We are only concerned with players who have a realistic chance to win MVP, not secondary players, so we plan to filter the dataset by requiring a certain number of games, minutes, and points so that the model won't need to be trained on insignificant data. Also, a crucial factor for MVP is team success, and while the dataset incorporates win percentage, it doesn't include the seed or rank they were in their respective conference. We plan to web scrape Basketball Reference or use the NBA API to pull data and combine the Kaggle dataset with this data to produce our final dataset for this project [5].

5 Group Members

Naman Tellakula, Vedanth Sathwik Toduru Madabushi

6 Ethical Implications

We are aware that inherent biases in the pre-trained models we plan to implement may exist and exacerbated through tuning our hyperparameters. Privacy concerns can also arise if sensitive data, especially NBA player data that is not publicly available, is used for training the models. Also, the dataset itself can have racial, ethnic, or other biases that could affect the model's output. Because of these potential ethical concerns, we monitor our dataset thoroughly, incorporate fairness and robustness into our evaluations, and adhere to ethical standards to promote safe and responsible development.

References

[1] Ishan Godbole, Sidhaarth Sredharan Murali, and Sowmya Kamath S. (2025). Nba mvp prediction and historical analysis using cross-era comparison approaches. Department of Information Technology, National Institute of Technology Karnataka, Surathkal, India.

[2] Armando Harlianto and Johan Setiawan. (2025). Forecasting the nba's most valuable player: a regression analysis approach. Information Systems Study, Universitas Multimedia Nusantara, Tangerang, Indonesia.

[3] Roberts Sunderhaft. (2024). Nba player season statistics with mvp win share. Kaggle. <https://www.kaggle.com/datasets/robertsunderhaft/nba-player-season-statistics-with-mvp-win-share>

[4] Tim West. (2018). The evolution of the mvp. Medium. <https://medium.com/prolific-interactive/the-evolution-of-the-mvp-5a2cbc9cca0e>

[5] NBA API Documentation Team. (n.d.). Nba-api documentation. <https://nba-apidocumentation.knowledgeowl.com/help>