Role-Driven Mixture of Experts for NFL Post-Snap Prediction

Shishir Aravindan
University of Toronto
shishir.aravindan@mail.utoronto.ca

Haashir Khan
University of Toronto
haashir.khan@mail.utoronto.ca

Myoungjae Kim
University of Toronto
myoungjae.kim@mail.utoronto.ca

Abstract

Teams in the NFL often analyze snaps, the exchange of the football between the center and the quarterback at the start of an offensive play, to determine strategies for their future games. However, predicting these post-snap outcomes is a complex task, as relying solely on human intuition can significantly impact strategic decision making. This paper aims to develop a deep learning model capable of accurately predicting the post-snap outcome based on the sequence of pre-snap data. By analyzing pre-snap sequences, the model will predict anticipated play outcomes, thus assessing the advantage for each team following the snap. This paper proposes utilizing a Mixture of Experts (MoE) model to capture the temporal and sequential dependencies offered by the data, inherently found in football plays. The model will use a Multilayer Perceptron (MLP) for the historical data, a Convolutional Neural Network (CNN) for the formational data, and a Recurrent Neural Network (RNN) for the pre-snap player movement data. The Gating Network, composed of another MLP, will be used to compute the final prediction. This architecture will break down the overall problem into subproblems, giving the model more flexibility by separately considering these major factors, which impact the outcome of the snap. Successful implementation of this model could provide NFL coaches and analysts with valuable insights, enhancing decision making, and optimizing game strategies.

1 Introduction

With the advent of advanced computational techniques, sport analytics has undergone a transformative evolution. Increasing availability of high-performance computing and breakthroughs in machine learning algorithms has enabled sophisticated predictive modeling across various domains, particularly in predicting outcomes in sports [1]. In professional football, organizations like the NFL have embraced deep learning to gain competitive insights. For instance, the league has partnered with Amazon Web Services (AWS) to deploy AI technologies to simulate in-game and practice scenarios to identify players at risk of injury [2]. These deep learning tools transform coaching strategies, enabling data-driven decisions that optimize performance and game planning.

Building on these promising AI capabilities, our research addresses a critical challenge in football strategy: predicting post-snap play outcomes with high precision. By integrating pre-snap contextual data—including player positioning, team formations, and historical performance metrics—we aim to develop a predictive model that can provide coaches and analysts with strategic insights.

37th Conference on Neural Information Processing Systems (NeurIPS 2024).

We propose a novel approach utilizing the Mixture of Experts (MoE) architecture [3], a sophisticated architecture capable of handling complex, diverse data domains. Our MoE approach divides the model into multiple "experts" that each specialize on different aspects of the problem, allowing more efficient learning and accurate predictions [3]. In our implementation, we decompose the prediction task across three critical domains: historical performance data, team formation characteristics, and sequential player interactions. This approach is well-suited for analyzing the complex and diverse nature of football data, where different types of plays and scenarios require distinct models of analysis.

By allowing specialized experts to analyze different aspects of the play, the proposed model leverages MoE architecture's inherent flexibility to generate more comprehensive and contextually aware predictions of post-snap outcomes.

2 Background and Related Work

2.1 Key Terms

Term	Definition
Down	An attempt by the offensive team to advance the ball at least ten yards within four consecutive attempts (downs).
Snap	The action where the football is passed from the center to the quarterback at the start of a play.
Formation	The arrangement of offensive or defensive players on the field.
Player Positioning	The placement of individual players on the field based on their role in the play.
Post Snap	The period after the ball is snapped.
Pre Snap	The time before the snap where players line up.
Play	A single instance of action on the field.
Run	A play where the ball carrier advances the ball by running.
Pass	A play where the quarterback throws the ball to a receiver to advance the ball downfield.
Frame	A specific period or snapshot of time during a play; 60 frames in a second.
Week	A specific week of tracking data collected of various NFL games.

Table 1: Key terms related to the NFL sport

2.2 Background

The National Football League (NFL) consists of 32 teams competing in a highly strategic, physical sport. Each game is segmented into a series of plays, with the offensive team's objective being to advance the ball toward the opponent's end zone to score. Plays are organized into a sequence of downs, and the offense has four attempts to progress at least ten yards or score directly.

The center of each play is the snap. The pre-snap formation sets the stage, influencing both offensive and defensive strategies. Once the ball is snapped, post-snap movements such as the quarterback's decisions, passing routes, and the offensive line's blocking schemes unfold, often in response to defensive pressures. These complex, fluid player movements, combined with other game dynamics, makes NFL in-game prediction particularly challenging.

Yet, the development of in-game tracking technology, along with advancements in machine learning and the decreasing cost of computational resources, has enabled more sophisticated analysis of the game. This has led to a growing body of research and practical applications for predictive modeling in the NFL.

2.3 Related Work

The field of NFL play prediction has seen significant advancements in both academic research and commercial applications. Existing work have explored various dimensions of post-snap prediction, particularly focusing on pass/rush prediction and post-snap receiver prediction. We draw insights from such studies to guide our methodology on feature engineering and model design.

A foundational study by Teich et al. in "NFL Play Prediction" [4] introduced a 'progress' metric that evaluates the success of a play based on down, yards gained, and yards to go. While simplistic, this metric provides valuable insights into feature engineering by emphasizing contextually relevant play dynamics. In their comparative analysis of machine learning models for pass/rush prediction, the authors found that Radial Basis Function(RBF) SVMs outperformed alternative models by more effectively capturing non-linear relationships in the data. Although we do not directly use their 'progress' metric, their work informs our approach by highlighting the need to: (1) address context-dependent input features, (2) model sequential data, and (3) select model architectures that capture subtle non-linear patterns in football play dynamics.

Further research for the same pass/rush prediction, such as this MIT thesis by Goyal [5], demonstrates the need for team-specific models, as they outperformed more generalized ones. This insight underscores the value of specialization in model design, guiding us to consider capturing team-specific characteristics and dynamics in both our feature engineering and model design.

Another dimension of specialization emerges in the work of NFL prediction [6], which highlights the power of an ensemble strategy. The authors employed a committee of machines (CoM)—specifically Artificial Neural Networks— for prediction. The study also introduces the use of statistical differentials, which are differences between comparable statistics(e.g. yards gained versus yards lost), to provide better contextualization of match-ups compared to raw statistics. Using sequential data similar to our use case, but from an earlier season, this study provides two key insights: First, it informs us on how to extract context-dependent features from the data thus creating more relevant/richer representations of the input data. Second, the finding that the CoM approach yields a higher predictive accuracy than a single model approach highlights the importance of adopting an ensemble strategy in designing our model.

Based on these insights from related work, we incorporate a Mixture-of-Experts (MoE) architecture in our model design. The MoE architecture dynamically selects relevant experts based on the input data, allowing each expert to specialize in a specific aspect of the task [3]. Given the multi-dimensional nature of the input data—pre-snap positions (sequential data), current game data, and historical performance—this role-driven approach offers an effective solution. Each expert is assigned a distinct role to process a subset of the input data, with each one predicting the binary outcome independently. The outputs are then combined in a static manner, ensuring that all experts contribute consistently to each prediction when making the final decision. This role-driven specialization design ensures balanced contributions from each input, enhancing overall model performance.

3 Data

Data	Description	# Features	# Rows
Game	Teams playing in each game.	9	136
Play	Information regarding plays for each game.	50	16124
Player	Player-level info from tracking data files.	7	1697
Players Play	Player stats for each game and play.	50	354727
Tracking	Player tracking data from weeks 1 to 9.	18	52740239

Table 2: Summary of dataset files

3.1 Dataset Overview

The NFL tracking dataset from the Big Data Bowl 2025 competition [7] provides an unprecedented granular view of game dynamics. The dataset comprises five interconnected data sources: game metadata, play-level information, player details, player-play statistics, and high-resolution tracking data. To get a better understanding of the data composition and scale, refer to Table 2.

The data captures a wide range of game-play contexts, with variations in player actions and outcomes across different game situations. This high-dimensional, player-centered data structure introduces modeling complexity, as players' roles and actions change dynamically across game situations.

3.2 Data Cleaning and Preprocessing

The data cleaning process focuses on handling inconsistencies and ensuring the data is standardized across the different game situations. Given the multi-dimensionality of the dataset, partitioning was done based on context (player formation, play sequence, game features), and each partition was preprocessed according to the expert handling that subset of the data.

A normalization technique was required to address inconsistencies in the *Tracking* dataset, where varying numbers of frames per player were observed across the pre-snap and post-snap sequence. To ensure consistency, we randomly sampled the same 10 frames for each player from the available frames in both the pre-snap sequences. This ensured that the temporal features had uniform shape.

3.3 Exploratory Data Analysis (EDA) & Data Insights

Existing literature emphasizes the importance of sequential pre-snap movement and spatial formation data for predicting game outcomes, prompting the inclusion of a dedicated expert for these features. Our EDA further refined feature selection. First, we confirmed the "home team advantage," with home teams showing a 53% win rate (Figure 1), highlighting the value of game context features. Second, to capture individual player performance, we used percentile-based features (e.g., passing yards for quarterbacks) instead of raw stats. These features, validated by their alignment with independent rankings (Figure 2) to official rankings [8], provide better context for modeling relative performance.

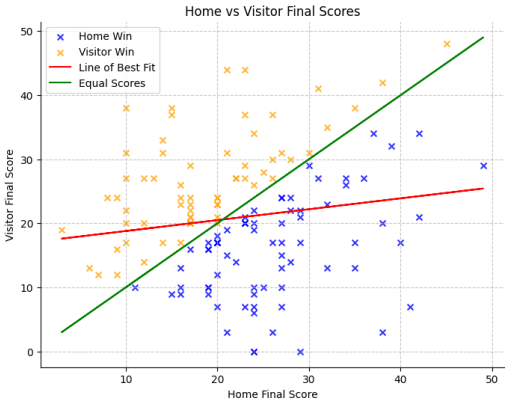

Figure 1: Home Team advantage

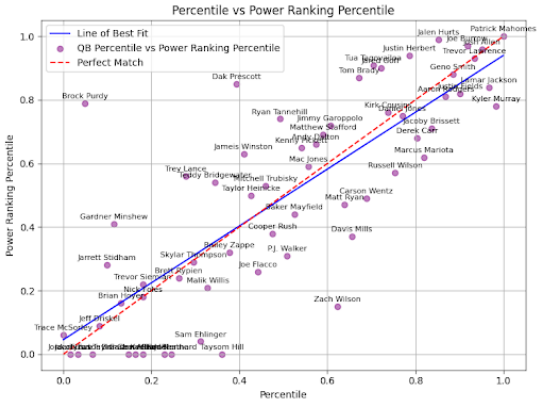

Figure 2: QB Offensive Output by Percentile

3.4 Data Engineering

Building upon the insights from the exploratory data analysis and preprocessing, we developed an ETL (Extract, Transform, Load) workflow to implement the feature engineering process.

The data was partitioned based on expert context, with each expert in the MoE architecture handling a specific data modality. For example, the Offensive Coordinator Expert(RNN) processes the pre-snap positional data of the offensive team. This data is treated as a sequence of 10 frames of 11 players. Relevant features are then engineered to capture the dynamics of this input modality. For instance,

we added a `timeToSnap` feature to provide temporal context. Finally, some features are normalized, such as converting the player's positional data (eg: `x`, `y`) relative to the quaterback.

A similar approach of data engineering is applied to the Formation Expert (CNN) and Game Context Expert (MLP), where data was partitioned, features identified and engineered, and normalization techniques applied for each respective modality.

4 Model (Github Link)

Given the complex nature of our data our proposed architecture is a Mixture of Experts (MoE) model designed to predict post-snap outcomes by leveraging spatio-temporal data and game context. The model includes an ensemble of specialized experts that process varied modalities of input data. The MoE architecture allows the prediction model to tailor its focus to the most relevant input features for a given prediction, ensuring both accuracy and interpretability.

4.1 Model Architecture Figure

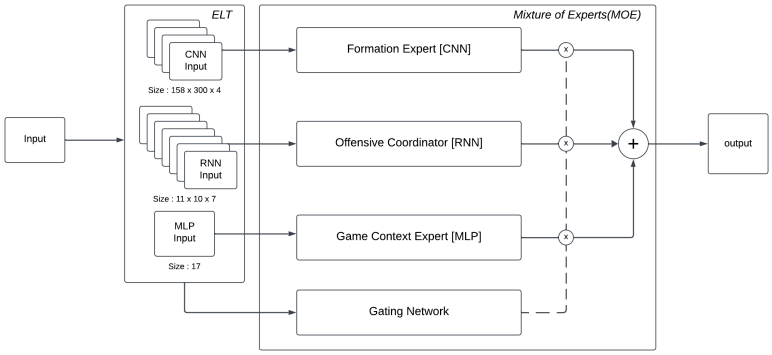

Figure 3: Mixture of Experts Proposed Architecture

4.2 Model Architecture

Each expert within the MoE architecture is designed to specialize in a specific subset of the input data. Hence they are trained independently by engineering features for their respective data modality. Below, we detail the design and rationale for each expert.

4.2.1 Offensive coordinator (RNN)

The input to the RNN model has the shape (11, 10, 7), where:

- **Input Dimension**: 11 represents the number of players (fixed) on the offensive team.
- **Sequence Length**: 10 frames representing pre-snap positional data tracked over time.
- **Feature Dimensionality**: 7 features tracked per player per time step (e.g., x-coordinate, y-coordinate, orientation, velocity, acceleration, etc.).

The model uses a stacked bidirectional LSTM network with two layers of 128 hidden units, capturing both past and future dependencies in the data. The bidirectional structure improves context understanding, while a dropout layer (rate 0.2) reduces over-fitting by randomly dropping connections. Finally, a dense output layer maps the features to a single value for binary classification.

4.2.2 Formation Expert (CNN)

The CNN will analyze the spatial formations of both teams, which will be represented in a $(300, 158, 4)$, where 300×158 is the spatial resolution of the field, with 4 input channels, representing the RGB and depth values of the input image. The image contains all player positions at the time of the snap.

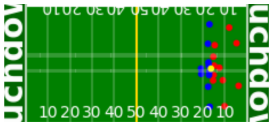

Figure 4: Example Input Image for a Pass Play

The architecture includes 3 convolutional layers with 32, 64, and 128 filters, each using a 3×3 kernel, batch normalization, ReLU activation, and max pooling. A 0.5 dropout layer is included to prevent over-fitting, increasing the models ability to generalize by forcing it to not rely on specific neurons. The convolution layers output a compact feature vector capturing spatial relationships, which is processed by 3 fully connected layers and sigmoid activation to generate the final prediction.

4.2.3 Game Context and Historical Performance Expert (MLP)

This expert will handle structured data, derived from the state of the game and historical factors that would impact the game. Inputs include both categorical and numerical features, such as the current down, the current quarter, yards to go, and player performance metrics. Categorical variables are encoded through embeddings, while numerical features are normalized.

The architecture consists of a feature vector with an input size of $(15,)$, followed by four fully connected layers; 3 hidden layers of size 50 and 1 output layer. The hidden layers will use ReLU activation to capture the non-linear relations, while the output layer uses sigmoid activation to determine the binary classification.

4.3 Gating network

The gating network in our MoE architecture uses an MLP to assign weights to experts [9]. The gating network takes the input and generates a set of weights, which decided how much each expert should contribute to the final prediction. These weights are then used to scale the output of each expert model. Finally, the weighted outputs are combined to produce the final prediction.

5 Results

Models	Offensive Coordinator (RNN)	Game context Expert (MLP)	Formation Expert (CNN)	Combined (MOE)
F1 score	0.6707	0.3925	0.5955	0.6216
Test Accuracy	0.6060	0.6243	0.6170	0.5789

The main metrics chosen to evaluate the individual experts and the overall MoE include their F1 score, accuracy. The F1 score is a measure of the harmonic mean of precision and recall. Precision measures the proportion of true positives out of all positive predictions made by the model, and recall measures how accurately the model identifies positives from the actual dataset labeled as positive. In our dataset there is a natural imbalance as the teams prefer passing compared to running[10]. Therefore, solely relying on accuracy to measure our model performance can lead to misleading conclusions.

The F1 score ensures that the model not only performs well overall but also performs effectively across both run and pass play types, especially in an imbalanced dataset. By incorporating the F1 score alongside accuracy, we can gain a more comprehensive understanding of the model's performance and evaluate its effectiveness in real-world scenarios.

The accuracy shows the overall consistency in making predictions; how accurate the model's predictions are. Therefore, without taking into account the balance of the dataset, a higher accuracy usually implies a stronger model capable of making general predictions.

6 Discussion

6.1 Improvements to Expert Models

The Offensive Coordinator model processes a 10-sequence of 7 features: x position, y position, orientation, speed, acceleration, direction, and angle of 11 players. Initially, the x positions ranged from 0 to 120 and the y positions from 0 to 53.3, which introduced sparsity that could negatively impact model predictions. To address this, we transformed the x and y positions to be relative to the quarterback's position. While this type of normalization is typically expected to improve accuracy, we did not observe a significant improvement in this case. Potential reason for this outcome could be the inconsistent order of players in the input. Since, the players in the snap frequently change, it is difficult to consistently align their order. This lack of alignment likely introduces noise into the input data, preventing the model from fully leveraging the normalized relative positions and other features.

During the training of a Formation expert, multiple hyperparameters were tuned, and additional layers were added to the model. For example, including a dropout layer likely reduced overfitting by regularizing the model, it forces the model to rely less on specific neurons. Hyperparameter tuning involved testing different kernel sizes, strides, and padding settings, ultimately landing on a 3x3 kernel size with a stride of 1 and no padding. These changes contributed to the model's high F1 score, indicating that these enhancements enabled the model to generalize more effectively.

For the Game Context Expert, grid search was used to tune hyperparameters such as the activation function, learning rate, batch size, and backpropagation method. Ultimately, the ReLU activation function, a learning rate of 0.001, and a batch size of 32 were selected as the best hyperparameters. However, the low F1 score indicates that while grid search successfully improved certain aspects of the model, it failed to address its difficulty in handling rarer patterns.

6.2 Baseline Models

The Offensive Coordinator(RNN) model demonstrated moderate performance on the test datasets, achieving an F1 score of approximately 0.75 and an accuracy of around 60%. These results indicate that the model generalized reasonably well without overfitting, as validation and test metrics align closely.

The Formation Expert(CNN) displayed similar performance to the RNN, achieving an F1 score of 76% and an accuracy of roughly 0.62. Its strong F1 score suggests that the model effectively captures spatial patterns within player formations. While this indicates that the CNN is well-suited for the data, its test accuracy leaves room for improvement, such as further tuning kernel sizes and other hyperparameters.

The Game Context Expert(MLP) achieved a moderate test accuracy of 62% but displayed a low F1 score despite having a relatively low test loss of 0.39. This suggests that the model struggles with imbalanced or underrepresented patterns, as evidenced by its difficulty in generalizing predictions. The F1 score highlights the model's challenge in handling rarer cases.

6.3 Mixture of Experts (MoE) Performance

The Mixture of Experts (MoE) model achieved an accuracy of 57% and F1 score of 0.69 on the test set. Initially, the model lacked normalization, which caused it to assign all weights to the best-performing expert. This was contrary to our goal, as we intended to incorporate predictions from different experts for a final decision. To address this issue, we implemented L1 normalization with a weight decay of 0.01. While normalization improved the integration of other experts, the model still leaned heavily on the best-performing expert, which is logical given the gating network's priority to maximize accuracy.

However, the results were unexpected. The MoE achieved lower accuracy than some of individual expert and performed worse than the MoE without normalization. The cause could be due to the complex nature of the problem itself. On many occasions, the individual experts produced conflicting predictions. While this was anticipated given the distinct inputs used by each expert, the

gating network in the MoE struggled to combine these conflicting predictions. The introduction of normalization likely forced the gating network to consider weaker experts, adding noise and ultimately degrading accuracy.

6.4 Comparison with Baselines

The accuracy and F1 score achieved by the MoE model were lower than those of the individual experts and the baseline models. This suggests that the MoE struggled to combine the strengths of the individual experts effectively. The gating network's difficulty in resolving conflicting predictions, coupled with the added complexity of normalization, likely contributed to these results.

7 Limitations

The main limitation of our MoE model is the lack of diversity among experts. The model heavily relied on the Game Context Expert, which provided the best accuracy, while other experts were underutilized. As a result, the model struggled to make accurate predictions when the Game Context Expert gave incorrect prediction. Despite regularization, the model continued to favor specific experts. To address this, a dynamic gating network is needed to adaptively select the most appropriate expert based on the input, rather than using static weights for all inputs.

8 Ethical Considerations

Although the model's design is intended for analytical purposes, it raises ethical concerns.

One key issue is the potential misuse of predictive models for unethical betting, as improved accuracy can foster harmful gambling behaviors, as highlighted by Binesh et al. [10].

To address this, we document the model's intended use, limitations, and impact to promote transparency and responsible use. These tools may also affect the sport's integrity by diminishing the role of human coaches and analysts, whose intuition and expertise are vital to the game. Over-reliance on technology risks eroding the human element central to its spirit.

Finally, data-related concerns are addressed by using player-consented data provided by the NFL via Kaggle, licensed under CC BY-NC-4.0. While we cannot share the data directly due to competition rules, it remains accessible to others under the same terms.

9 Conclusion

The main goal of this study was to develop a predictive model with a deep learning aspect to correctly predict a binary outcome for NFL plays by leveraging spatio-temporal data and game context. Three specialized experts were used; Offensive Coordinator (RNN) for sequential patterns in the data, Formation Expert (CNN) for analyzing team formations for each play, and Game Context Expert (MLP) for contextual analysis. These models were evaluated individually and combined in a Mixture of Experts (MoE) framework.

The MoE approach demonstrated strong predictive performance, with high F1 scores for both run and pass plays. The RNN expert, while effective in capturing sequential data patterns, was limited by its handling of contextual information. Similarly, The CNN excelled at identifying spatial patterns but required significant computational resources, making it prone to overfitting. The MLP performed well in contextual analysis but was limited by the imbalanced dataset.

In conclusion, By integrating these models, the MoE framework leveraged their strengths and mitigated individual weaknesses, resulting in improved overall performance. This study shows the importance of integrating spatio-temporal information and higher-level context in making accurate predictions. Future work may further explore class imbalance handling, increasing computational efficiency, and further refine contextual insight inclusion to increase the predictive accuracy for a wide variety of game scenarios.

References

[1] Nitin Singh. Sport analytics: A review. *The International Technology Management Review*, 9:64–69, 01 2020.

[2] NFL. Nfl and amazon web services expand partnership to further shape the future of football, 09 2024.

[3] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017.

[4] Brendan Teich, Roman Lutz, and Valentin Kassarnig. Nfl play prediction, 2016.

[5] Udgam Goyal. Leveraging machine learning to predict playcalling tendencies in the nfl. Master's thesis, MIT, 12 2019.

[6] John A. David, R. Drew Pasteur, M. Saif Ahmad, and Michael C. Janning. Nfl prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports*, 7(2), 2011.

[7] Michael Lopez, Thompson Bliss, Ally Blake, Paul Mooney, and Addison Howard. Nfl big data bowl 2025, 2024.

[8] Marc Sessler. Nfl qb index: Ranking all 68 starting quarterbacks from the 2022 nfl season, 02 2023.

[9] Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton. Adaptive mixture of local expert. *Neural Computation*, 3:78–88, 02 1991.

[10] Fantasy Football Today. Nfl run/pass ratios - 2022, 2022. Accessed: 2023-10-15.