National Football League Big Data Bowl 2025

NFL

Use Pre-snap Behavior Analysis

Pre-Snap to Post-Snap Analysis: Insights, Predictions, and Visualizations

Leveraging Data, LLMs, and Machine Learning for Enhanced Game Understanding by

- Noor Alsabahi
- Mahmood Rahimi
- David Van Ginneken

Introduction and Goal

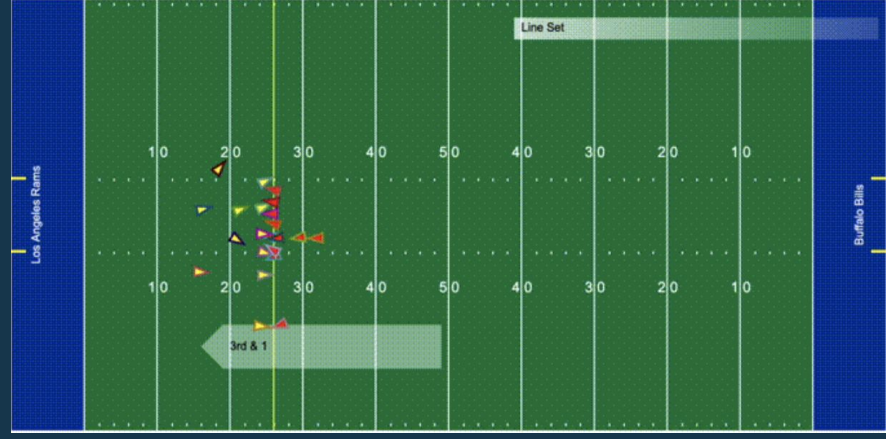
Objective

- Provide actionable insights on offensive and defensive strategies.
- Predict post-snap outcomes based on pre-snap behaviors.
- Enable coaches, analysts, and players to understand tendencies through a user-friendly website.

Key Components

1. Data Analysis and Visualization
2. LLM-Driven Descriptive Insights
3. Machine Learning for Yardage Prediction
4. Game Playback for Historical Comparison

Unified Web Platform



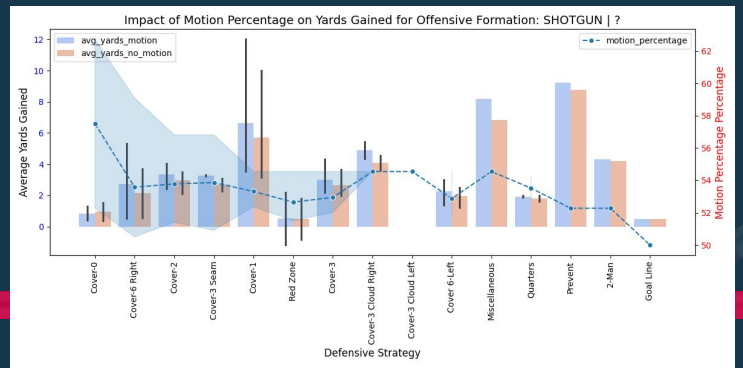
Features

- Integrated services for seamless analysis, predictions, and visualizations.
- Intuitive UI for team and strategy selection and running full analysis.
- Interactive 2D game play using tracking data for analysis.

Flow

1. Select teams and strategies.
2. Analyze pre-snap tendencies.
3. View descriptive insights or run predictions.
4. Watch historical playback for similar scenarios.

Claude



Transforming NFL Play Data for Strategic Insights

Data Analysis

Objective:

- Prepare and enhance the NFL dataset to extract actionable features.

Steps Taken:

- **Data Merging:**
  - Integrated data across multiple NFL datasets for comprehensive analysis.
- **Feature Engineering:**
  Created key indicators:
  - Pre-snap motion presence (`any_motion`).
  - Post-snap actions (`is_pass`, `is_run`).
  - Time bucketing

Insights Generated:

- Aggregated data to extract offensive and defensive strategies insights
- Established foundational metrics such as:
  - Average yardage gained.
  - Run/Pass post-snap percentages.
  - Motion pre-snap percentages.
  - Formation utilization percentages.

Key Metrics & Strategic Insights Derived

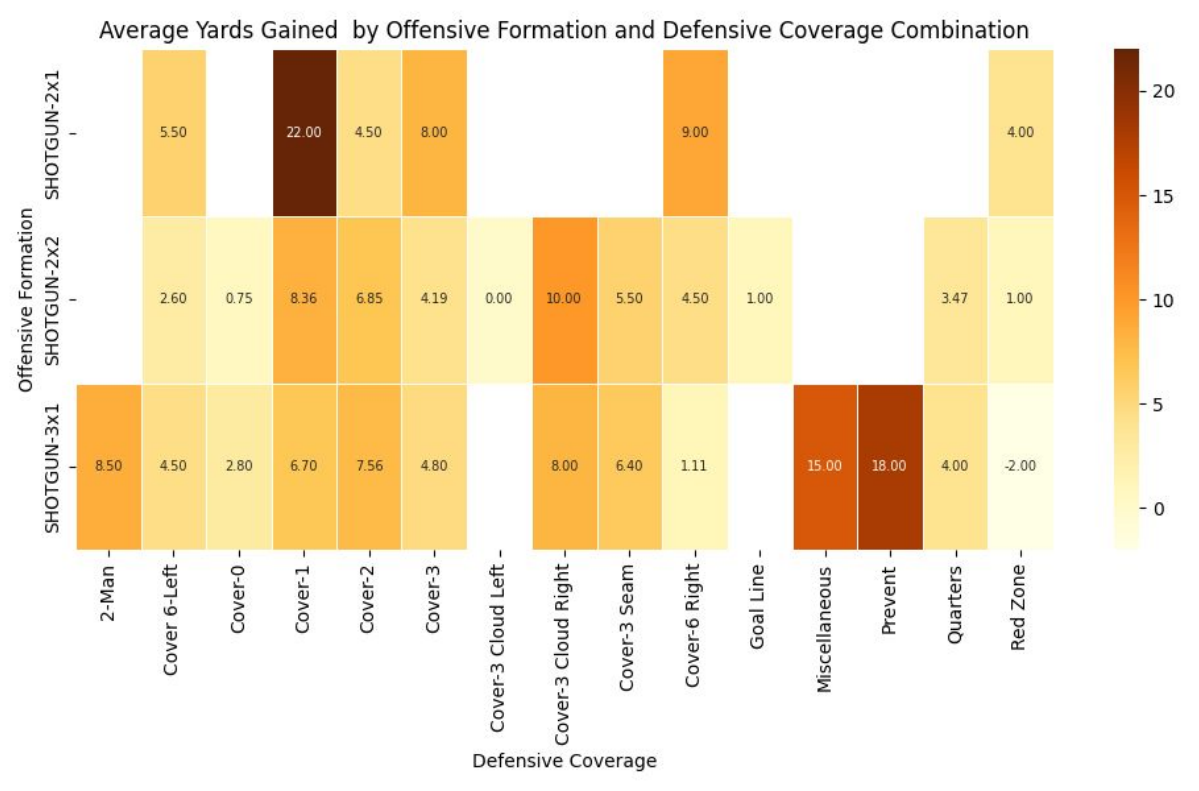
Aggregation and Strategy Analysis

- **Data Aggregation:**
  - Grouped by **offensive formation**, **receiver alignment**, and **defensive coverage**.
- **Calculated Metrics:**
  - **Average Yards Gained**: Overall and with/without motion.
  - **Success Rate**: Percentage of successful plays.
  - **Pass/Run Frequencies & Rates**: Quantifying offensive tactics.
  - **Motion Utilization**: Percentage of plays with pre-snap motion.
  - **Formation Percentage**: Utilization of various formations.
- **Sorting Criteria:**
  - Highlighted **success rate** and **average yards gained** to uncover strategic advantages.

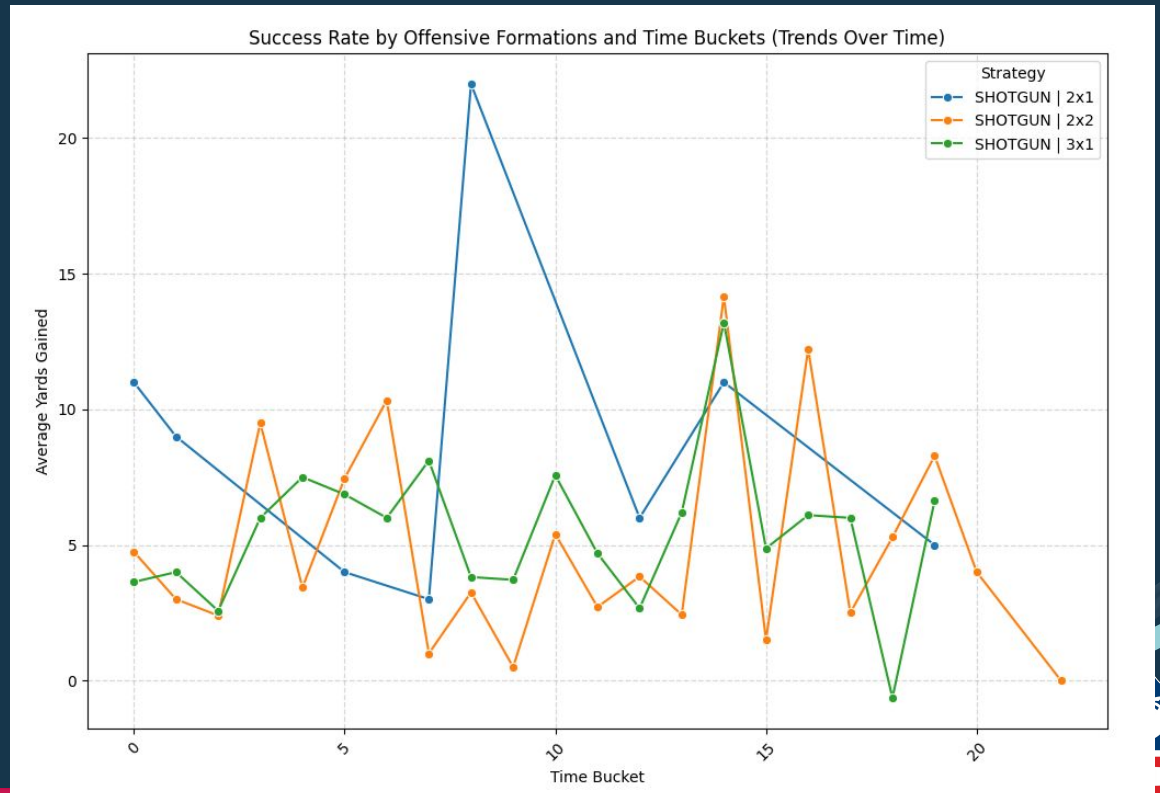
Visualize Strategic Insights

Visualizations for Strategy Effectiveness Conducted

Heatmap: Shows average yards gained by offensive and defensive strategies.



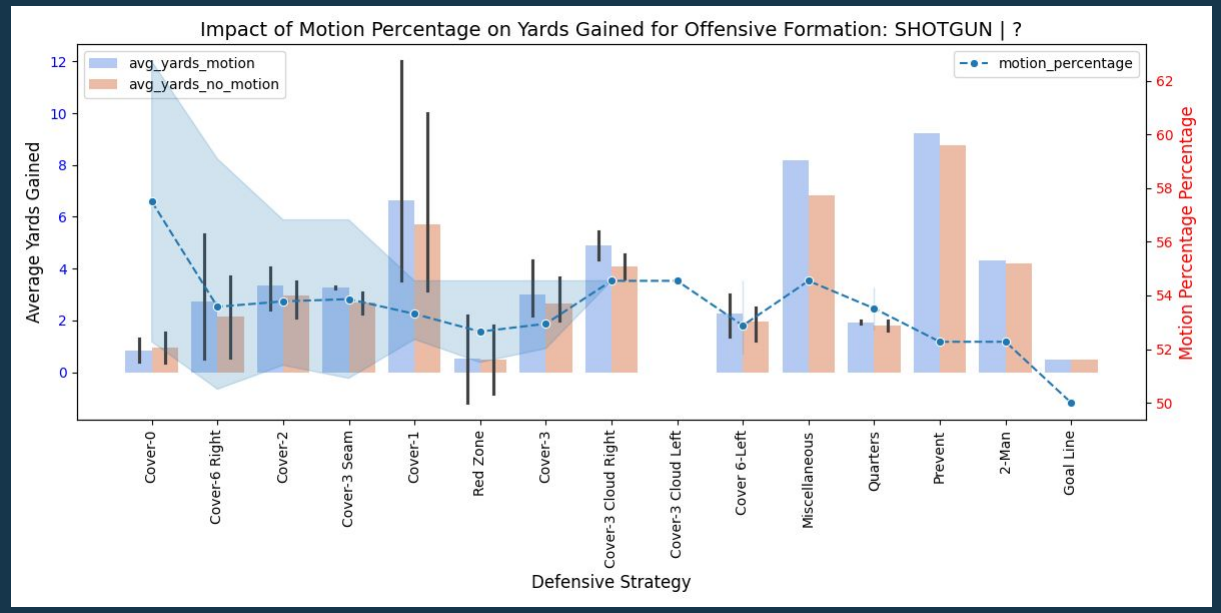
Line Plot: Captures the temporal impact of specific strategies on effectiveness.



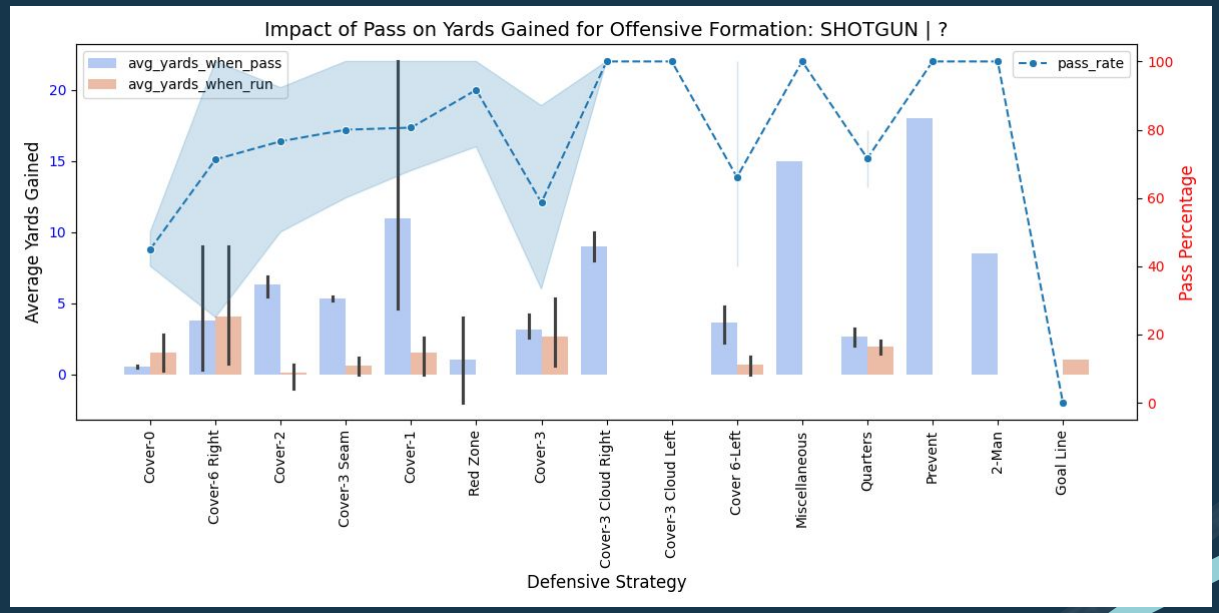
Visualize Strategic Insights

Visualizations for Strategy Effectiveness Conducted

Bar Plots: Pre-snap motion effects on offensive/defensive effectiveness.



Bar Plots: Run/Pass post-snap effects on play effectiveness.



LLM-Driven Descriptive Insights

Translate Complex Data into Actionable Insights:

- **Challenge**: Raw data is often too granular and complex for immediate interpretation.
- **Solution**: LLMs work on **aggregated data**, which is pre-processed to provide high-level summaries and insights. This makes it easier to derive actionable conclusions that decision-makers can use.

Provide Comparative Analysis:

- **Comparative Analysis**: By working with aggregated data, LLMs can compare various strategies or formations across multiple games or scenarios.
- **Use Case**: Comparing performance between different offensive strategies in football (e.g., plays with pre-snap motion vs. plays without motion).

Automated and Scalable Reporting:

- **Efficiency**: LLMs generate insights from **aggregated data**, saving time compared to manually analyzing individual data points.
- **Scale**: Analyze large volumes of data at once to uncover trends and patterns that would be difficult to spot manually.

How it Works

- **Data Aggregation**:
  - Raw data is pre-processed into **aggregated summaries** (e.g., average yardage, play type distribution) for easier analysis.
- **Prompt Engineering**:
  - The aggregated data is fed into a well-crafted **LLM prompt** that asks for a specific comparative or descriptive analysis.
- **Model Response**:
  - The LLM generates **human-readable insights**, summarizing trends and providing actionable recommendations based on the aggregated data.
- **Actionable Insights**:
  - Insights are interpreted and used for strategic decision-making, refining strategies or identifying patterns for future actions.

LLM Advanced Analysis Example - Part 1

NFL Game Data Analysis: Pittsburgh Steelers vs. New England Patriots

Offensive Strategy Analysis

Formation and Alignment Effectiveness

Based on the provided data, the most effective offensive formations and alignments for the Pittsburgh Steelers are:

- SHOTGUN | 3x1: This formation appears frequently and shows good versatility against various defensive coverages.
- SINGLEBACK | 3x1: Another common formation that yields positive results, especially against Cover-1 and Cover-3 defenses.

Passing vs. Running Strategy

While specific pass and run frequencies are not provided, we can infer from the formations that the Steelers rely heavily on passing plays, especially from the SHOTGUN formation. To improve their chances of success, the Steelers should:

1. Maintain a balanced attack by incorporating more running plays from the SINGLEBACK formation to keep the defense guessing.
2. Utilize pre-snap motion more frequently, as plays with motion tend to gain more yards on average.

Defensive Strategy Analysis

The New England Patriots employ various defensive coverage schemes, with Cover-1, Cover-2, and Cover-3 being the most common. To counter these strategies, the Steelers should:

- Against Cover-1: Exploit man-to-man matchups using quick slants and crossing routes.
- Against Cover-2: Target the deep middle of the field and use underneath routes to exploit zones.
- Against Cover-3: Utilize flood concepts and attack the seams between zones.

Motion Analysis

While specific motion statistics are not provided, it's generally accepted that pre-snap motion can create mismatches and confusion for the defense. The Steelers should increase their use of motion, especially when facing complex coverage schemes like Quarters or Cover-3 Cloud.

LLM Advanced Analysis
Example - Part 2

Time-Based Strategy Analysis

Early Game (Time Buckets 0-4)

In the early stages of the game, the Steelers should focus on:

- Utilizing SHOTGUN | 2x2 and SHOTGUN | 3x1 formations, which show good early-game success.
- Mixing in SINGLEBACK | 2x2 and SINGLEBACK | 3x1 formations to establish the run threat.

Mid-Game (Time Buckets 5-9)

As the game progresses, the Steelers should:

- Increase the use of SHOTGUN | 3x1 formation, which shows high yardage gains (up to 23 yards) against Cover-2.
- Be cautious with SINGLEBACK | 3x1 against Cover-2, as it resulted in a 7-yard loss in one instance.

Late Game (Time Buckets 10+)

In the latter stages of the game, the Steelers should:

- Rely heavily on SHOTGUN formations, which consistently gain positive yardage.
- Use SINGLEBACK | 3x1 against Cover-1, which yielded an 18-yard gain in time bucket 10.

Comparative Analysis: Steelers Offense vs. Patriots Defense

To improve their chances of success against the Patriots' defensive strategies, the Steelers should:

1. Exploit the Patriots' Cover-3 with SHOTGUN | 2x2 formations, which have shown success with gains of up to 12 yards.
2. Use SINGLEBACK | 2x2 against Cover-2, which resulted in a 14-yard gain in time bucket 10.
3. Be prepared to adjust to the Patriots' varied coverage schemes by having a mix of plays ready from both SHOTGUN and SINGLEBACK formations.
4. Utilize EMPTY | 3x2 formations sparingly, as they have not shown significant success in the provided data.

Actionable Insights

1. Increase the frequency of pre-snap motion to create defensive mismatches and gain an edge in yardage.
2. Focus on SHOTGUN | 3x1 and SINGLEBACK | 3x1 formations, adapting the choice based on the defensive coverage observed.
3. Develop a strong package of plays from the SHOTGUN | 2x2 formation for early game situations.
4. Practice exploiting Cover-2 defenses with deep middle routes from SHOTGUN formations.
5. Prepare specific play sequences for late-game scenarios, emphasizing SHOTGUN formations with quick-hitting passes.

By implementing these strategies and remaining adaptable to the Patriots' defensive adjustments, the Pittsburgh Steelers can maximize their offensive effectiveness and improve their chances of success in the matchup.

Machine Learning Model

Goal

- Predict average yardage gained based on inputs like team, strategy, and formation.

Model Overview

- Input Features: Offensive and defensive strategies, alignments, pre-snap motions.
- Output: Predicted average yardage.
- Validation: Achieved a high R^2 score, indicating robust predictive capability.

Feature Engineering & Data Preparation

Objective: Develop a predictive model to estimate yardage based on offensive and defensive teams, their strategies, game time, and field position.

Steps:

1. **Data Cleaning**: Removed outliers to ensure high-quality input data.
2. **Stratified Splitting**: Dataset split into training, validation, and testing sets based on possessionTeam to maintain team-specific distribution.
3. **Feature Engineering**:
    - Transformed gameClock and quarter into a unified _time_remaining_in_game feature.
    - Calculated _distance_to_goal based on yardlineNumber and playDirection.
    - Standardized numerical features and applied one-hot encoding for categorical variables:
        - Teams (possessionTeam, defensiveTeam)
        - Strategies (offenseFormation, receiverAlignment, pff_passCoverage)
4. **Pipeline Implementation**:
    - Automated feature preparation using scalable pipelines.
    - Ensured consistency across training, validation, and testing datasets.

Model Design & Training

Model Structure:

- **Input Layer:** Features engineered from game data.
- **Hidden Layers:**
  - 32 neurons, Tanh activation, L2 regularization, and Batch Normalization.
  - Dropout layer (50%) for regularization.
  - 16 neurons in the second layer with similar configuration.
- **Output Layer:** Single neuron for yardsGained prediction with linear activation.

Optimizations:

- The learning rate scheduler, to optimize convergence and prevent stagnation during training.
- Dropout and Regularization (L2 and dropout) to prevent overfitting.
- Early stopping and learning rate reduction to fine-tune training.
- Batch Normalization: Stabilizes and accelerates training by normalizing activations.

Training Details:

- Optimizer: Adam
- Loss Function: Mean Squared Error (MSE)
- Achieved Best RMSE: ~5

Comprehensive Model Evaluation: Pearson Correlation & RMSE

In addition to evaluating the model's performance with **Root Mean Square Error (RMSE)**, we used **Pearson correlation** to assess the relationship between predicted and actual values. The results are as follows:

- **Pearson Correlation Coefficient**: A value of **0.1432** indicates a **weak positive correlation** between the predicted and actual values. This suggests that while there is some alignment between the predictions and the true values, the model's performance can be further improved.
- **Statistical Significance**: The **p-value** of **0.0000** is **statistically significant** (p-value < 0.05), indicating that the observed correlation is unlikely to be due to random chance and that the relationship between predictions and actual values is meaningful.
- **Opportunity for Improvement**: While the correlation is statistically significant, the relatively low correlation value highlights the need for further refinement of the model. The model's predictive power can be enhanced as more data becomes available.

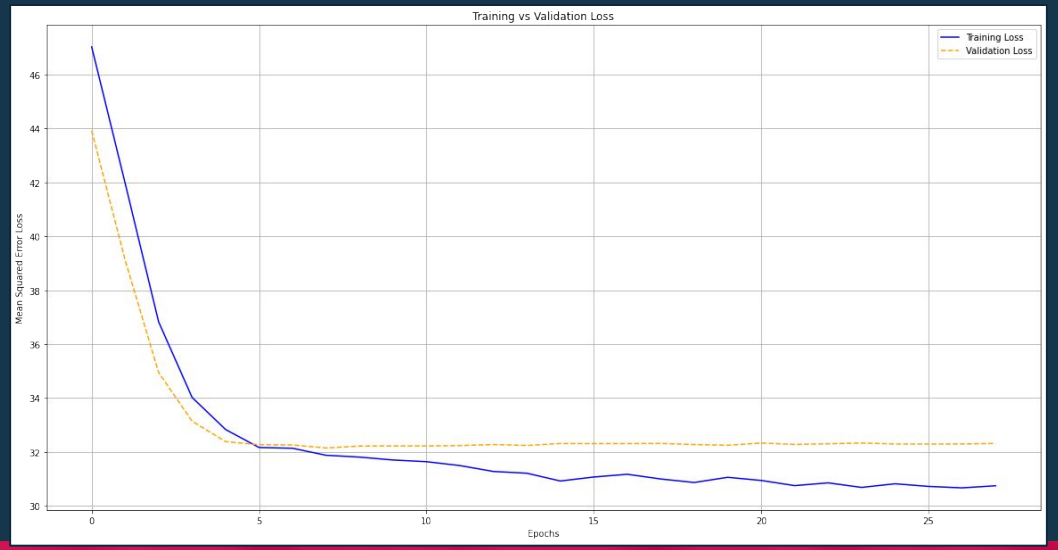
Limitations and Future Improvements

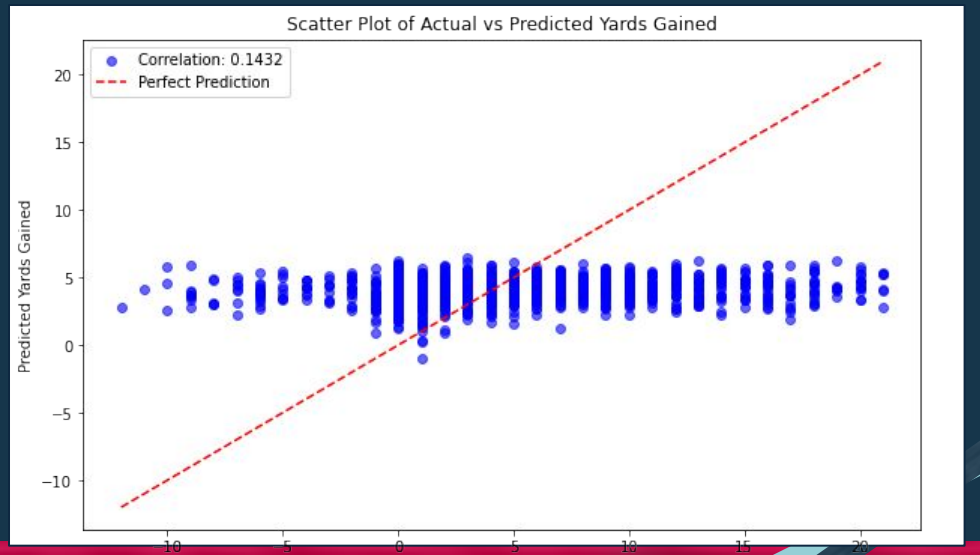
Current Limitations:

- **Data Quantity:** Limited dataset size restricts further RMSE improvement.
- **Generalization:** Model performance could improve with more diverse data.

Future Prospects:

- Access to larger datasets for better generalization and model robustness.



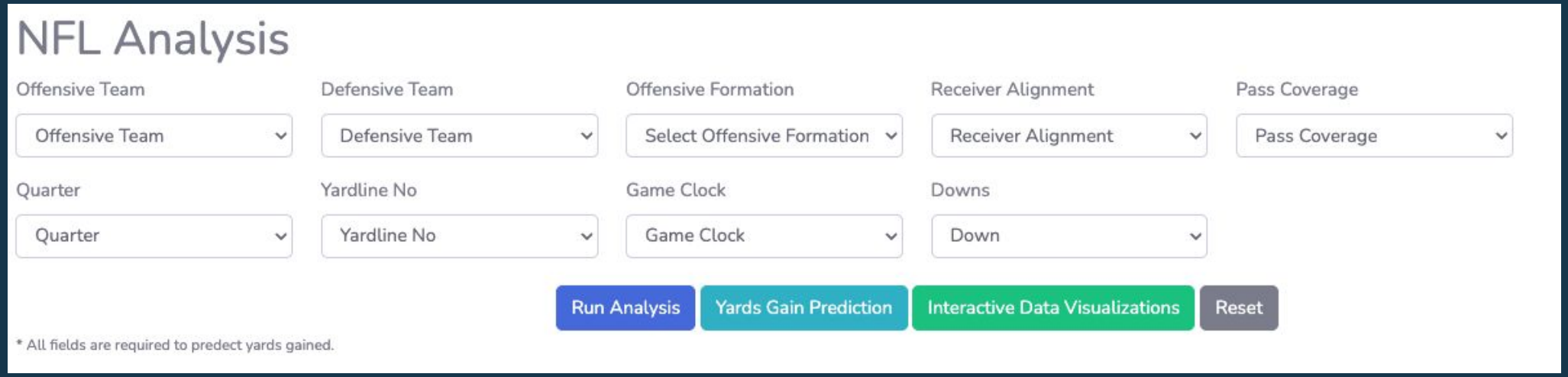


Getting Started : Data Analysis

User driven analysis

- **Select desired offensive and defensive alignments**
- **Review predicted yardage gain/loss**

Example



Game Analysis via 2D Visualization

Interactive Playback

- Replay the a play of the game using all the data. A full football field is shown.
- HTML5 Canvas is used to fully play the tracking data by play.
- Canvas also animates ball pass and play outcome.
- All players have unique colors and can be further identified by clicking on player name.

Use Case

- Coaches are able to see a 2D replay of the game without the distraction of a real game play and cameras moving around. They are able to pause and reset game play and control speed.

LA VS. BUF

Play Number: 236

Quarter: 1

Down: 3

yards To Go: 1

(10:03) J.Allen pass short right to G.Davis for 26 yards, TOUCHDOWN.

Bobby Wagner

Position: ILB

Aaron Donald

Position: DT

Troy Hill

Position: CB

Jalen Ramsey

Position: CB

A'Shawn Robinson

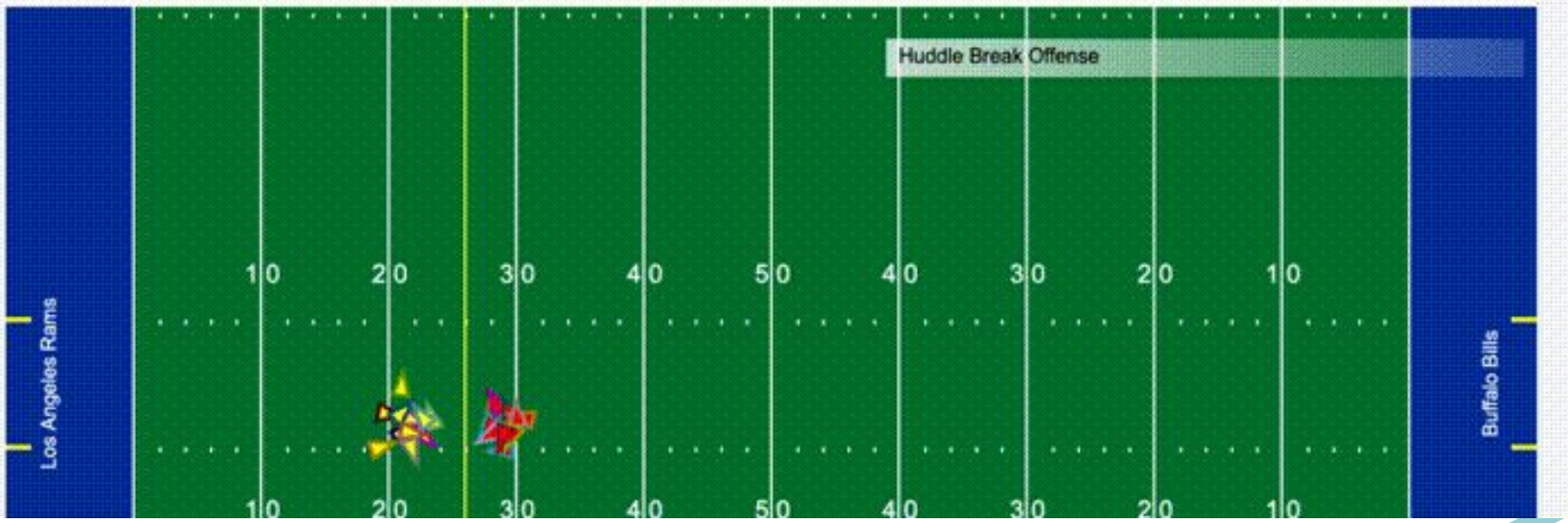
Position: DT

Taylor Rapp

Position: SS

Pause Reset

Low Speed Medium Speed High Speed



Rodger Saffold

Position: G

Mitch Morse

Position: C

Dion Dawkins

Position: T

Josh Allen

Position: QB

Dawson Knox

Position: TE

Tommy Sweeney

Position: TE