Explainable Defense Coverage Classification in NFL Games using Deep Neural Networks

Huan Song[1], Mohamad Al Jazaery[1], Haibo Ding[1], Lin Lee Cheong[1], Jonathan Jung[2],
Mike Band[2], Michael Chi[2] and Tom Bliss[2]
[1] Amazon ML Solutions Lab, [2] NFL Next Gen Stats (NGS)

1. Introduction

Machine learning (ML)-powered football analytics has received considerable interest in recent years [1, 2, 3, 4, 5, 6, 7], with majority of existing analytic measures centered around offense strategies and performances [5, 6]. In contrast, the defensive side of the game has received relatively less attention and development. At the core of understanding and analyzing any defensive strategy is the **coverage scheme**, *i.e.*, the rules and responsibilities of each defender tasked with stopping the pass. Classifying the coverage scheme for every pass play provides insights and new understanding to the football game to teams, broadcasters and fans alike. The preferences of play callers become apparent through coverage scheme data, such as Bill Belichick using Cover 1 at a top 5 rate in five consecutive seasons. Coverage scheme classification also allows deeper understanding on how respective coaches and teams continuously adjust their strategies based on their opponent's strengths. For example, the Packers and Chiefs both faced significantly more man coverage through the first 11 weeks of the 2022 season than they did in 2021 after both teams traded away their leading receivers during the offseason (Davante Adams & Tyreek Hill, respectively). Finally, coverage scheme classification enables the development of new defensive-oriented analytics such as uniqueness of coverages [18]. In 2020, Brandon Staley designed the most unique set of coverages for the Rams while the fired Gregg William's was the least unique.

Manual identification of these coverages on a per-play basis is both laborious and difficult as it requires football specialists to carefully inspect the game footage. Thus, there is a need for an automated coverage classification model to effectively and efficiently scale to reduce cost and turn-around time. This coverage classification model also needs to address the inherent ambiguity around the deployed coverage schemes that can be difficult to grasp even for expert reviewers. For example, the defensive coaching staff will often disguise their coverages to mislead the quarterback. It is thus important to develop model explanation method to facilitate the understanding of what the machine learning model utilized to classify these coverages and arrived at a given conclusion. Figure 1 below shows the location of all offensive and defensive players at the start of an example play (left) and in the middle of the same play (right). The model showed relatively low confidence in its coverage classification on this play, with the top two predictions (Cover 3 Zone & Cover 1 Man) falling under 50 percent. The play action fake and the defenders' reactions to it along with the route distribution made it harder for the model to determine whether it was man or zone coverage.

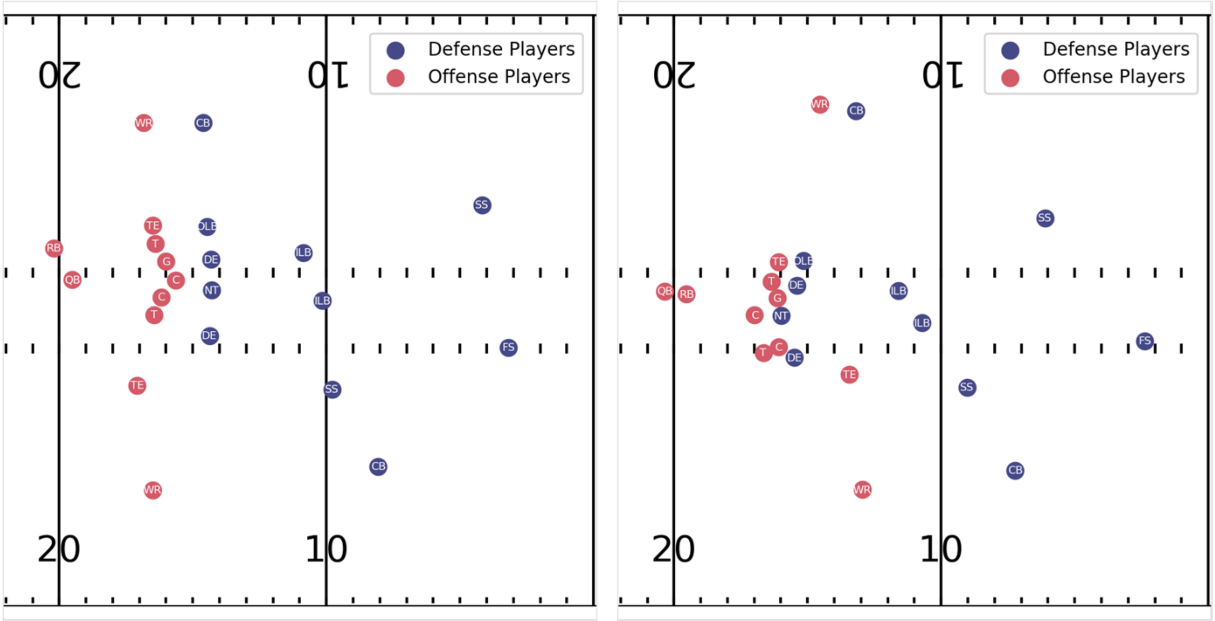


Figure 1: Example of an ambiguous play that shows the complexity of the task. Left, at the start of the play, and right in the middle of the play. Full list of player acronyms is in Appendix.

To the best of our knowledge, ML-based coverage classification has not been fully studied. Previous efforts from [8] dabbled on this topic by adapting the convolutional neural network (CNN)-based Kaggle Zoo winning solution of the 2020 Big Data Bowl [17], but ignored the temporal progress of the play. Based on our analysis, this approach struggled in achieving sufficient accuracy needed for productionization and reduction of manual review. Production readiness is defined here as achieving >95% accuracy in identifying man versus zone-type plays, as well as ability to determine plays that require further expert reviewing. In this paper, we present a novel deep learning approach that significantly outperforms [8] for automatic coverage classification. Raw sensor data comprised of location, speed and acceleration is collected for every player and utilized as inputs into an automatic coverage classification pipeline. We baseline using the published CNN-based model [8, 17] as well as the improved versions with incorporated long short-term memory (LSTM) component. We find that our proposed addition of attention layers results in improved classification accuracy, as these layers enables the model to learn to focus on specific aspects of a play. Further performance gain is achieved by applying label smoothing to tackle the inherent challenges in distinguishing the intricate coverage schemes, and model ensemble to bootstrap decisions from multiple independently trained base models. Finally, we incorporate model explanations via play embedding analysis and gradient-based approaches that provide confidence that the notoriously opaque deep learning model correctly captures football knowledge, and aligns with human experts' understanding. These model explanations also help speed up visual review processes, and bring additional insights about defense coverage schemes.

This remainder of the paper is organized as follows: we review the related work on tracking-based football analytics, coverage classification, and model explanation in Section 2. In Section 3, we present our coverage modeling and model explanation approaches. In Section 4, we describe our evaluation results on coverage classification and model explanation results. In Section 5, we conclude by summarizing our approach and results, and outlining our planned future works.

2. Related work

2.1 Tracking-based football analytics

Football tracking data contains rich information of the game dynamics including the player and ball location, speed, acceleration in real-time. This enriched large-scale data has attracted multiple in-depth studies to analyze team and specific player's performance, including trajectory prediction [7], quarterback evaluation [5, 6], pass inference penalty prediction [9], receiver openness and expected gain prediction [3], and run vs. pass prediction [4]. Other published works focused on expanding the analytics capability, either with additional data sources or improved model architecture design. In [1], the authors demonstrated the importance of incorporating charting annotations with tracking data. Authors in [2] focused on developing more advanced architecture components for feature representation learning to tackle the variable duration problem of events and the ordering problem of players. In [10], a graph neural network was developed to better capture the player interactions and their fast progression over time.

2.2 Football coverage classification

Despite the criticality of analyzing and understanding the defensive strategies, it has only been investigated in a few works so far. Dutta et.al. [11] developed an *unsupervised* learning approach to group each player's pass coverage into the high-level man vs. zone categories. The approach from [12] focused on team-wide defense coverages, but is based only on vision data. The most relevant work to ours is [8], where B. Baldwin developed a convolutional neural network to identify eight defensive coverage schemes. However, only a single frame from each play is utilized for the identification. The temporal progression of the player location and interactions contains critical information about the coverage scheme, and relying on the static features from certain frame could significantly limit the predictive power. In this paper, we design and describe new architectural components that tackle the temporal modeling challenge and beyond, leading to a performant classification model.

2.3 Model explanation for sports analytics

Although deep neural network models have achieved remarkable results in various sport analytics problems, its black-box nature prohibits interpretation of how it came to the conclusion. The explainability, however, is critical in 1) extracting additional insights on the data and predictive task, 2) verifying that the model correctly captured the related sport knowledge, and 3) indicating when human experts should be involved in-the-loop to resolve any prediction issues. The explainability of sports analytics models was studied only recently. In [13], interpretable decision tree-based models were developed along with neural network models for football pass vs. rush prediction to study how much accuracy of DNNs they can capture. A case study on outcome prediction of volleyball matches was conducted in [14] that utilized different explanation approaches including Boolean Rule Column Generation, ProtoDash, and SHAP (SHapley Additive exPlanations). For baseball predictions, [15] utilized Shapley values to get both local feature importance and global feature importance for batter vs. pitcher plate appearance (PA). [16] leveraged LIME (Local Interpretable Model-agnostic Explanations) for NBA gameplay predictions that discovered insights leading to the success of a given NBA team. These works focused on the explanation of high-level statistical features such as player historical performances. Our work in this paper provides comprehensive understanding on both the global level that discovers important samples of interest for manual review, and for the first time, on the instance level that uncovers the leading evidences on the fine-grained play tracking data.

3. Task Definition, Data, and Methods

3.2 Task Definition

We define the defensive coverage classification problem as a multi-class classification task, with three types of man coverage (where each defensive player covers certain offensive player) and five types of zone coverage (each defensive player covers a certain area on the field). These eight classes are visually depicted in Figure 2 below: Cover 0 Man, Cover 1 Man, Cover 2 Man, Cover 2 Zone, Cover 3 Zone, Cover 4 Zone, Cover 6 Zone and Prevent (also zone coverage). Multitude of information over time must be accounted for to properly identify the correct coverage, including the way defenders lined up before the snap, the adjustments to offensive player movement once the ball is snapped, coverage disguises and even blown coverage assignments.

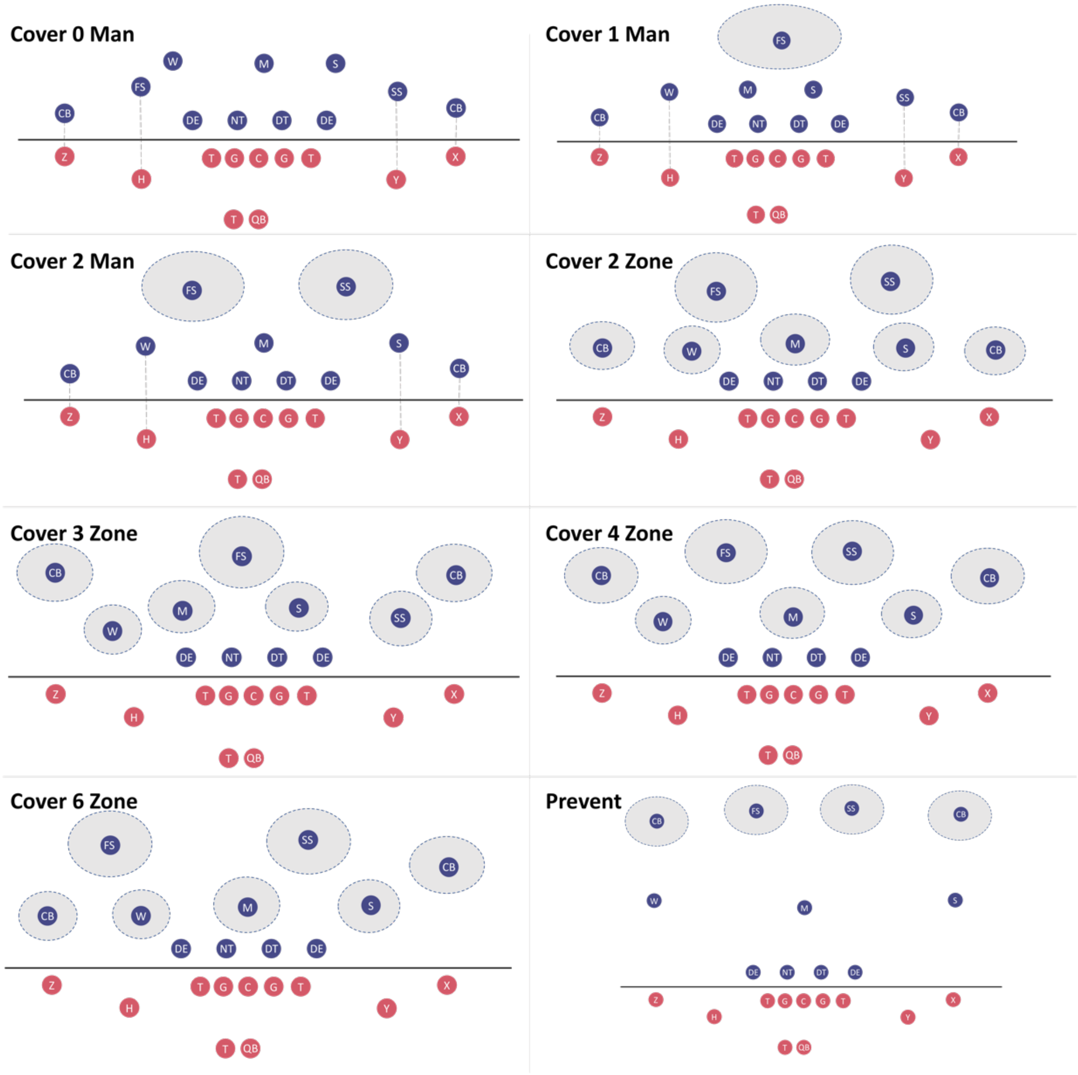


Figure 2. Defensive coverage types considered in our classification task. Circles in blue are the defensive players laid out in a particular type of coverage; circles in red are the offensive players. Full list of player acronyms is in Appendix.

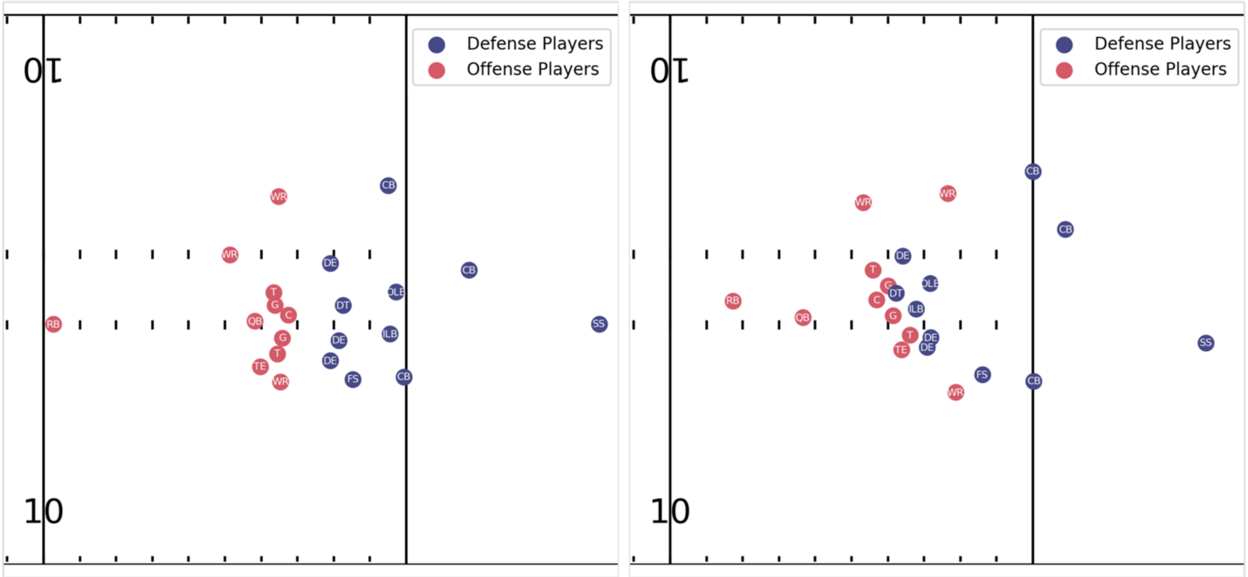


Figure 3. Player tracking data illustration on the snapshots of the 1st frame (left) and the 10th frame (right) of a Cover 1 Man play. A human reviewer visually inspects the entire play, taking into account multitude of interactions and positions, before making the final determination that this is a Cover 1 Man play. Full list of player acronyms is in Appendix.

The complexity and time-dependency of correctly identifying a coverage is illustrated in Figure 3, which shows two timed snapshots for a Cover 1 Man play. The offensive players are depicted in red, and the defensive players in blue. The letter within the blue and red circles denotes the player position on the field. In order to correctly determine the coverage as Cover 1 Man, the human reviewer or the model needs to account for the 1) interaction between the wide receivers (WR) and cornerbacks (CB), 2) interaction between the running back (RB) and linebackers (OLB, ILB), and 3) the location of the safety in the middle of the field (SS) as a single-high safety patrolling the deep middle area, over the duration of the play.

3.2 Data

Game tracking data is captured at 10 samples per second, including the player location, speed, acceleration and orientation. This is available for every player and every play from 2018 to 2021 by NFL's Next Gen Stats. We utilize 2018-2020 seasons data for model training and validation, and 2021 season data for model evaluation. Each season consists of around 17000 plays. Initial data cleaning was applied to remove noise introduced by sensor errors. For model training, we utilize the tracking data and the manually annotated coverage labels.

We plot the coverage class distribution and its change over seasons in Figure 4. The data shows unbalanced distribution over the classes where Cover 1 Man and Cover 3 Zone are dominant and Prevent class is in the minority. This is to be expected: Cover 1 Man and Cover 3 Zone are the two base coverages in modern football and Prevent coverage is a situational play call mostly saved for end of regulation situations. Additionally, the distribution over the seasons highlights the fact that the Man-type coverages popularity is generally decreasing season by season from 2018 to 2021 compared to the Zone-type coverages.

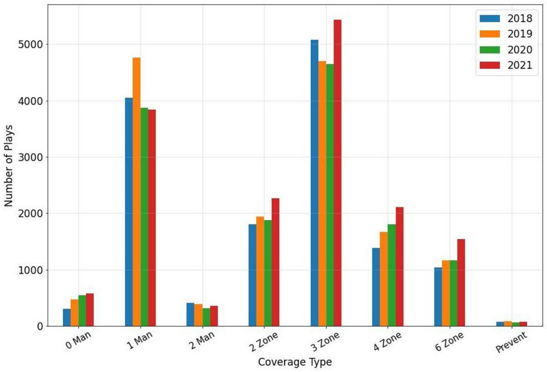


Figure 4. Coverage class distribution over 2018-2021 seasons.

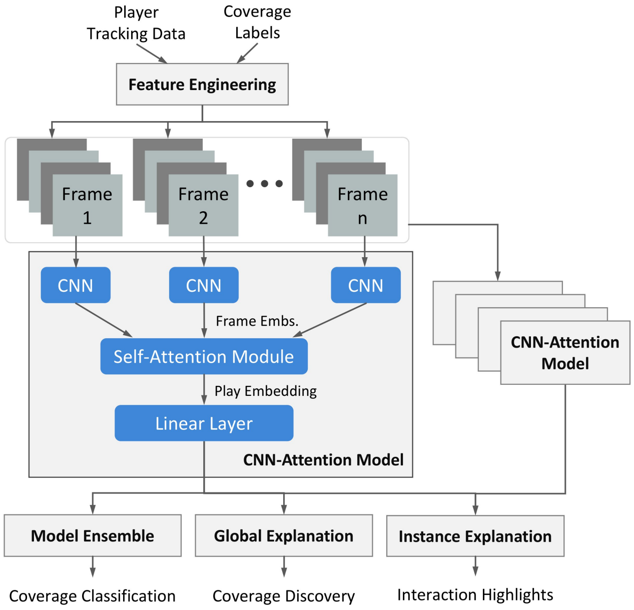


Figure 5. The explainable coverage classification framework, starting with inputs from the top of the sketch. Detailed information about the model is in Section 3.5 and about the explanations in Section 3.6.

3.3 Explainable Coverage Classification Framework

Figure 5 illustrates our overall modeling framework, with the input of player tracking data and coverage labels starting at the top of the figure. Given the input, we first conduct feature engineering to construct the player pair-wise relative features similar to [8], and then utilize convolutional neural network (CNN) to model the complex player interactions similar to the Kaggle Zoo solution [17]. Unlike [8] and [17], we apply a self-attention module that learns to aggregate the frame embeddings to focus on the most critical time steps, and an ensemble model that pools the decisions made by each model individually. The pooled decision is the output coverage classification. In addition, we develop a comprehensive model explanation method based on the learned play embeddings to provide both global and instance explanations. Global explanation utilizes embedding analysis to uncover potentially problematic plays for manual review, whereas instance explanation utilizes gradient-based CNN explanation to highlight most critical player interactions leading up to the identified coverage. In the next sections, we will describe details in the feature and data engineering (Section 3.4), CNN-attention model architecture (Section 3.5), and model explanation methods (Section 3.6).

3.4 Data and Feature Engineering

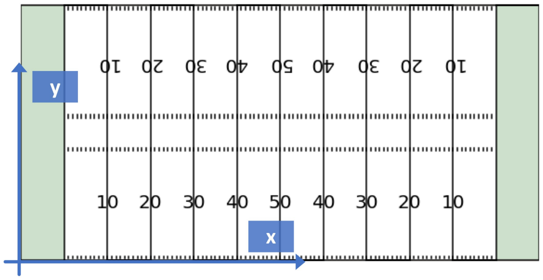


Figure 6. Data processing x, y definition shown on football field. Player raw features including the location, speed, acceleration etc. are decomposed onto the two axes.

```
Define start_event as "ball_snap",
end_events as [“pass_forward”, “qb_sack”,
“qb_strip_sack”, “qb_spike”, “tackle”, “pass_shovel”],
L as the length of play window containing frames
between start_event and any end_event.
IF L < 15 THEN
    pass
ELSE IF 15<=L<=45 THEN
    use the entire play window
ELSE
    use the first 45 out of the L frames
END IF
```

Algorithm 1. Play trimming algorithm.

We perform similar data processing steps as described in [8] and [17] that included decomposing raw features into x and y axes (as defined in Figure 6), unifying all play directions to left-to-right, and augmenting the y-axis location during training using random flipping with a 0.5 probability. We highlight key differences to [8] and [17] that we implement to improve the model’s performance:

- **Full offensive players.** [8] limited the feature engineering to 5 non-quarterback offensive players. We expand to full non-quarterback offensive players of 10 to maximize the input information. This provides the flexibility to let the model learn to capture the most important signals for coverage classification.
- **Play trimming.** We utilize a sequence of frames in the play to make the prediction, whereas both [8] and [17] were based on a single frame. As such, the duration of the play needs to be taken into account. Since the play lengths vary dramatically, we perform trimming of longer plays to focus on the first several seconds that contain the most important coverage indicators. The detailed trimming logic is described in Algorithm 1.
- **Temporal downsampling.** Due to the incorporation of full offensive players and additional frames from the play, the size of the input tensor to the model increases significantly. To reduce the memory footprint and make both training and inference more efficient, we experimented temporal downsampling of the play with different factors. We found downsampling by a factor of 2 (reducing to 5 frames per second) did not reduce the classification performance, and utilize it for all experiments in this paper.

After the raw data has been processed, we perform feature engineering to construct the play feature sequence as the input for model digestion. For a given frame, our representation is inspired by the Zoo model from 2020 Big Data Bowl Kaggle solution [17]: we construct an “image” for each time step with the defensive players at the rows and offensive players at the columns. The “pixel” of the “image” thus represents the features for the intersecting pair of players. Different from [17], we extract a sequence of the frame representations, which effectively generates a mini-“video” to characterize the play.

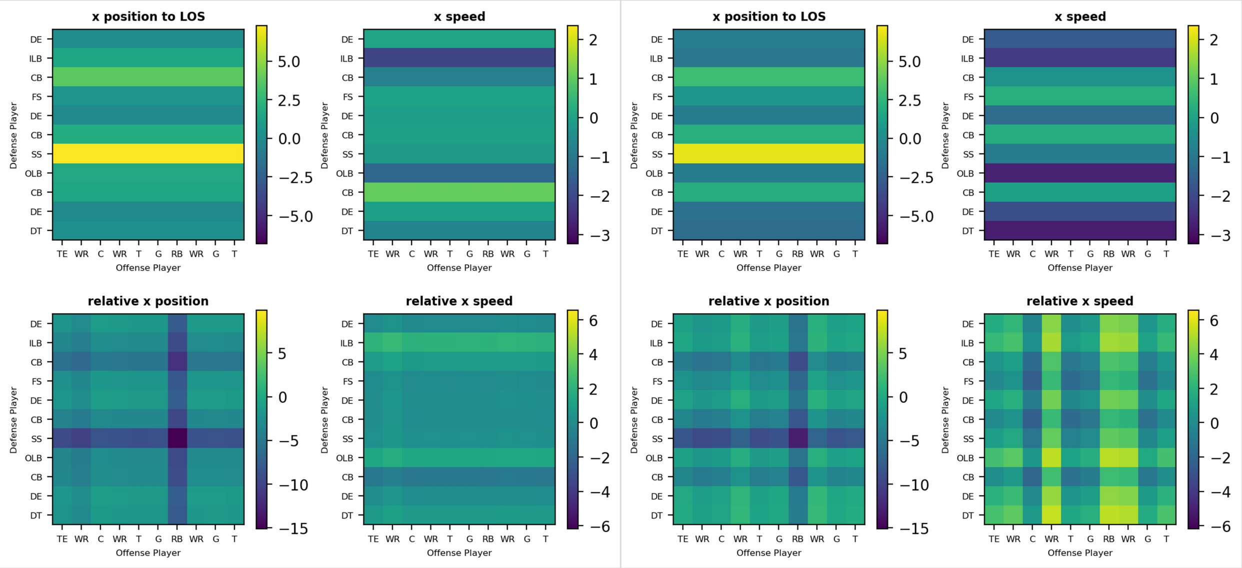


Figure 7. Illustration of the example features for the 1st frame (left) and the 10th frame (right) in correspondence to Figure 3. The x axis definition is given in Figure 6. Player acronyms are the same as in Figure 3 and the full list is in Appendix.

Figure 7 visualizes how the features evolve over time in correspondence to the two snapshots given in Figure 3. For visual clarity, we only show four features out of all the ones we extracted: “x position to LOS (line of scrimmage)” and “x speed” for defenders, which capture their location and speed on the horizontal direction of the play field; “relative x position” and “relative x speed” for the interacting defensive and offensive player pair, where the feature value is reflected at the “pixel”. The pixel color encodes the value according to the color-bar. Notice how the features progress over time as players move: for example, at 10th frame on the “relative x speed” feature, the 3 wide receivers (WR) columns have generally larger values, indicating the aggressing movements. On the other hand, on the “relative x position” feature, their intersecting “pixels” with SS and 3 CBs have relatively smaller values, indicating the close proximity these players got into. Comparatively, reading from the “x position to LOS” feature, large values for the SS and 3 CBs confirm their locations on the field.

Altogether, we construct the following two sets of features: 1) defender features consisting of the defender position, speed, acceleration and orientation, on x and y axis that corresponds to the horizontal and vertical direction of the field; 2) defender-offense relative features consisting of the same attributes but calculated as the difference between the defensive and offensive players. Aside from the player movement features, we also experimented with incorporating game contextual information including the down, yards to endzone, yards to go, number of pass rushers and running routes etc. These extra features did not show clear improvement of the coverage classification performance and was thus removed from the productionized pipeline. We conjecture that the rich tracking data inherently cover game and play information for the model and the context did not provide additional perspectives.

3.5 Coverage Classification Model

We develop an ensemble CNN-attention model that utilizes the features constructed in Section 3.4 for coverage classification. We describe the key architectural designs that are important for performant modeling in the next few subsections.

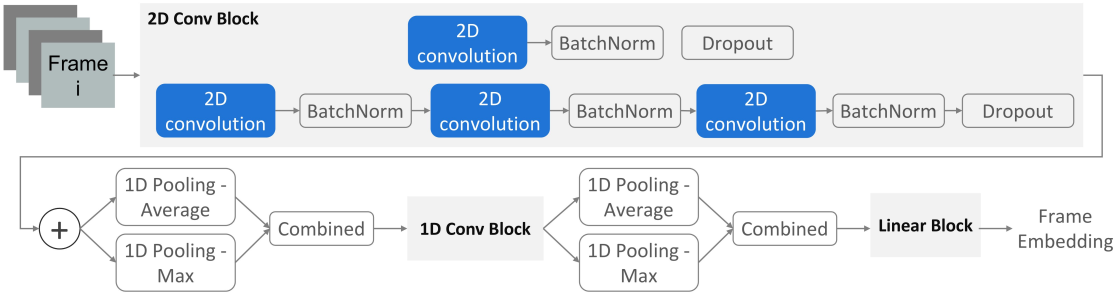


Figure 8. Diagram of the convolutional module

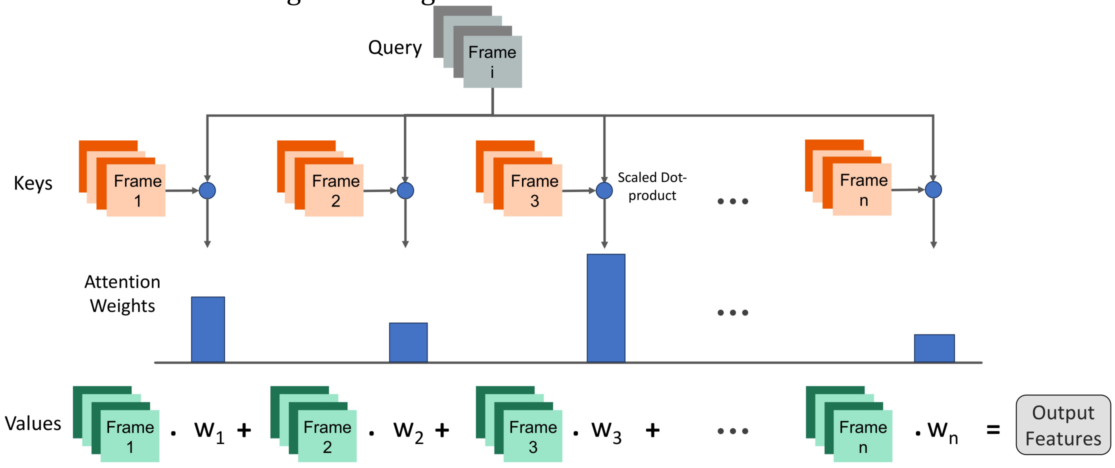


Figure 9. Self-attention mechanism for temporal modeling.

3.5.1 CNN module

The "image" feature construction as in Section 3.4 facilitated the modeling of each play frame through a CNN. Figure 8 shows the internal structure of our CNN: we modified the convolutional (Conv) block utilized by the Zoo solution [17] with a branching structure that is comprised of a shallow 1-layer CNN and a deep 3-layer CNN. Batch normalization is utilized after each convolution layer and dropout is applied at the end of the block. An important detail on the convolution layer is the internal 1x1 kernel: having the convolutional kernel looking at each player pair individually ensures that the model is invariant to the player ordering. In data processing, the ordering needs to be consistent over time for a play. For simplicity, we order the players based on their NFL ID for all play samples.

After the 2D Conv Block, pooling is applied along the offense axis ("image" columns). The results are then fed into a one-dimensional (1D) Conv Block composed of a similar structure as 2D Conv Block, but with 1D convolutional layers. Following [17], we utilize a weighted combination of average and max pooling with the weights of 0.7 and 0.3. We experimented with modified weights but the modifications did not provide any performance improvement. At the end of the CNN module is a linear block that consists of 3 fully connected layers with batch normalization and dropout in between. We obtain the frame embeddings as the output of the CNN module.

3.5.2 Temporal modeling

Once the ball is snapped, a play takes only a few seconds to complete. Within the short period, the fast-progressing, rich temporal dynamics contain key indicators to identify the coverage. The ML model needs to not only aggregate the information contained in individual frames, but also capture the correlations among the frames and potentially weigh them differently. We design a self-attention module [21] for the temporal modeling and compare it with a more conventional, bidirectional LSTM approach (quantitative comparison in Section 4.1). High-level illustration of the self-attention module is given in Figure 5, where the self-attention module is stacked on top of the frame embeddings learned from the CNN. The learned attention embeddings as the output are then averaged to obtain the embedding of the whole play. Finally, a fully connected layer is connected to determine the coverage class of the play.

We illustrate the internal structure of the self-attention module in Figure 9, where the attention weights are calculated as the scaled dot-product between each query frame and every key frames. The weights are then used in the linear combination of the value frames to compute the frame representations. Specifically,

$$MultiHead(Q, K, V) = Concat(head_1, head_1, \cdots, head_h)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

where K, V, Q are frame embeddings learned from the constructed "image" feature, W_i^Q, W_i^K, W_i^V are the layer weights, and h is the number of attention heads.

3.5.3 Model ensemble and label smoothing

As described in Section 3.2, the 8 coverage schemes have an imbalanced distribution: for example, Cover 1 Man and Cover 3 Zone are frequently utilized while Prevent and Cover 2 Man are rare. In addition, we identified adjustments in more specific coverage calls that can lead to ambiguity among the 8 general coverage classes for both manual charting and model classification. The coverage imbalance and ambiguity make the clear separation among coverages challenging.

We utilize model ensemble to tackle these challenges during model training. We experimented with the following ensemble methods: fusion, voting, and gradient boosting, along with different number of base models. In fusion, all base models are jointly trained with the training loss calculated from the averaged output of all base models. Voting differs from fusion by fitting base models independently, and averaging their outputs only during inference. For gradient boosting, the base models are trained sequentially, where the training target is associated with outputs from previously fitted base models. Our study shows that in fact, the more straightforward voting method achieves the best classification result and the 5-model ensemble works the best. In the voting-based ensemble, each base model has the same CNN-attention architecture and is trained independently from different random seeds. The final classification takes the average over the outputs from all base models.

We further incorporate label smoothing into the cross-entropy loss to handle the label ambiguity. The idea is to encourage the model to adapt to the inherent coverage ambiguity instead of overfitting to potentially biased annotations. To smooth the labels, in the loss calculation, the original one-hot class distribution is combined with a small amount of uniform class distribution to introduce uncertainty. For example, Cover 3 Zone label is modified as 90% probability of 3-Zone and equal probabilities of anything else. Denote the original one-hot encoded label vector for sample x as $y_x^{one-hot}$ and the number of classes as $K = 8$. Label smoothing is calculated as,

$$y_x^{label-smooth} = (1 - \alpha) y_x^{one-hot} + \alpha / K$$

where α is the tunable weighting parameter to control the smoothing strength. $y_x^{label-smooth}$ is then used in the cross-entropy loss calculation.

3.6 Model Explanations

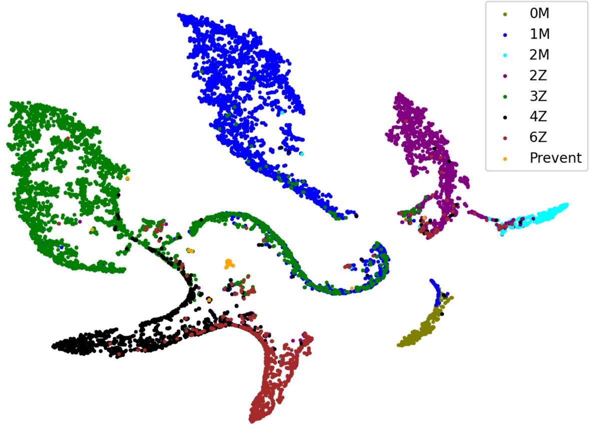


Figure 10. Global explanation: t-SNE embeddings of a downsampled subset of 2018-2020 season training plays. The plays are color-encoded according to the ground-truth annotation shown in the legend. The legends are shortened class labels of the original 8 classes as depicted in task definition, with M representing Man and Z representing Zone, and the word Cover removed.

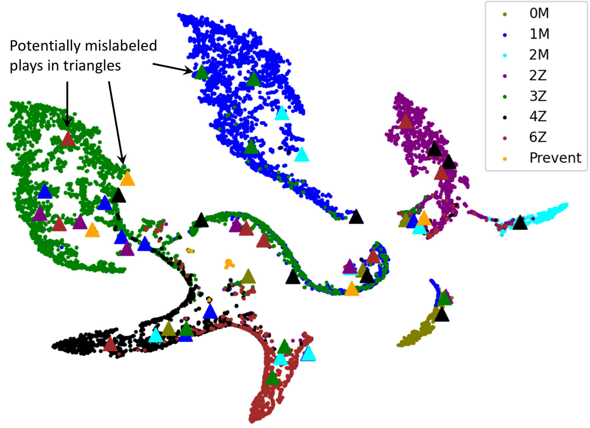


Figure 11. Potentially mislabeled plays highlighted on the t-SNE embeddings. The top-ranked identifications by the KNN algorithm are shown with triangles. The color encodes ground-truth coverage annotation.

The black-box nature of deep neural networks prohibits the interpretation of how it determines the coverage scheme from tracking data. Our analysis reveals the inherent challenges in ensuring that the model captures the football knowledge, and reviewing the model's decision under coverage ambiguities and wrong classifications. To tackle this, we develop a two-stage, top-down model explanation approach.

The first stage analyzes the learned play embeddings from the coverage classification model to discover any patterns that require manual review. We utilize t-distributed stochastic neighbor embedding (t-SNE) [26] and experimented with the parameters including perplexity, number of iterations and the random seed to extract stable 2D projections. To reduce visual clutter, we perform stratified sampling to analyze a subset of all training data that consists of around 9000 plays. The projected 2D embeddings are visualized in Figure 10. We find that the majority of each coverage scheme are well separated, demonstrating the classification capability gained by the model. However, we highlight two important patterns that need further investigation: 1) a small number of plays deviate significantly from their respective coverage cluster. This could be attributed to mislabels of the coverage or high degree of coverage ambiguity. 2) among certain coverages, there is significant overlapping of plays. For example, we identify a long, curved cluster consisting of a mixture of Cover 1 Man and Cover 3 Zone plays (blue and green samples) and the cluster deviates from the main clusters of both types. This could entail inherent ambiguity that can exist between these two coverage concepts and specific adjustments on play calls that are not accounted for in the general ground-truth labeling. To effectively extract the example plays associated with these patterns for manual review, we utilize basic outlier detection and unsupervised clustering methods. The detailed methods and our findings from manual review are described in Section 4.2.1.

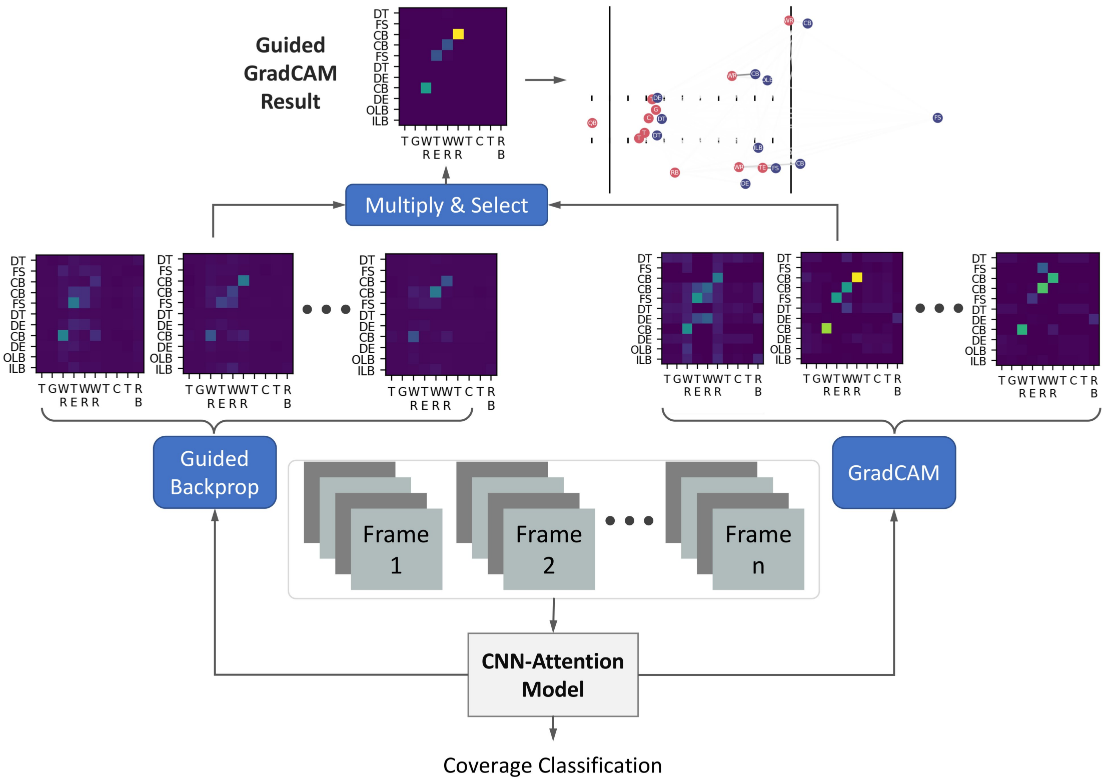


Figure 12. Instance explanation: we utilized Guided GradCAM algorithm to extract the highlighted pixels and mapped them back to football field where the line thickness corresponds to the player interaction strength.

To shed lights on model's decision and speed up the manual review on individual plays, we develop the second stage of instance explanation. It zooms into the individual play of interest, and extracts frame-by-frame player interaction highlights that contribute the most to the identified coverage scheme. This is achieved through Guided GradCAM algorithm [22] and the extraction process is illustrated in Figure 12. Starting from the coverage classification score obtained by the model (bottom of the figure), the algorithm consists of two steps. The first step (left branch in Figure 12) uses Guided Backpropagation [27] to extract the salient pixels of the input image that activate the neurons. These highlights are class-agnostic, general contributing features. The second step (right branch in Figure 12) uses GradCAM to back-propagate the coverage score using the gradients to localize class-discriminative pixels. Note that we utilize the feature maps at the output of the 2D Conv Block (as in Figure 8) to extract the GradCAM result. Results from these two steps are then element-multiplied and we select frame with the highest activation as the most critical time step for explanation. Considering the multiple base models (not shown in Figure 12 for conciseness) used in the ensemble, we also select the base model that outputs the highest activation. The explanation result is coverage-discriminative, pixel-level highlights on the transformed "image" feature as in Figure 7. As the final step to illustrate the highlights intuitively, we map them back on the football field and visualize the corresponding player interactions. The line thickness annotates the interaction strength. The detailed results on example plays are shown in Section 4.2.2.

4. Metrics and results

In this section, we describe the experimental metrics and results for our explainable coverage classification model. We first introduce the quantitative experiment setup and performance comparison to baseline models (Section 4.1). Next, we provide results from our model explanation methods and the insights discovered by them (Section 4.2).

4.1 Quantitative evaluation

Table 1. Best model and training parameters from hyperparameter optimization.

CNN output dimensionality	Learning rate	Weight decay	Label smoothing weight	Dropout rate for fully connected layers	Dropout rate for convolutional layers	Number of heads in self-attention module
128	0.0054	0.0005	0.07	0.3	0.2	16

As mentioned in Section 3.2, we utilize 2018-2020 seasons data for model training and validation, and 2021 season data to for quantitative evaluation. We performed a 5-fold cross-validation to select the best model during training. We apply the Adam optimizer with weight decay and perform hyperparameter optimization to select the best settings on multiple model architecture and training parameters. The best parameters are shown in Table 1.

To evaluate the model performance, we computed the coverage accuracy, F1 score, top-2 accuracy and accuracy of the man vs. zone task. The CNN-based Zoo model used in [8] is the most relevant for coverage classification and we used it as the baseline. In addition, we consider improved versions of the baseline that incorporate the temporal modeling components for comparative study: a CNN-LSTM model that utilizes a bi-directional LSTM to perform the temporal modeling, and a single CNN-attention model that is used as the backbone of our model, but without the ensemble and label smoothing components. We obtain the performance results from 5 runs with different random seeds and report the average and standard deviation measures. The results are shown in Table 2.

Table 2. Quantitative evaluation of the coverage classification model in comparison with the baseline and improved versions of it.

Model	**Test acc. 8 coverages (%)**	**Top-2 acc. 8 coverages (%)**	**F1 score 8 coverages**	**Test acc. Man vs. Zone (%)**
Baseline: Zoo model	68.8±0.4	87.7±0.1	65.8±0.4	88.4±0.4
CNN-LSTM	86.5±0.1	93.9±0.1	84.9±0.2	94.6±0.2
CNN-attention	87.7±0.2	94.7±0.2	85.9±0.2	94.6±0.2
Ours: ensemble of 5 CNN-attention models	**88.9±0.1**	**97.6±0.1**	**87.4±0.2**	**95.4±0.1**

We observe that incorporation of the temporal modeling module significantly improves the baseline Zoo model that was based on a single frame. Compared to the strong baseline of CNN-LSTM model, our proposed modeling components including the self-attention module, model ensemble and labeling smoothing combined provide significant performance improvement. The final model is performant as demonstrated by the evaluation measures. In addition, we identify very high top-2 accuracy and significant gap to the top-1 accuracy. This can be attributed to the coverage ambiguity: when the top classification is incorrect, the 2nd guess often matches human annotation.

4.2 Model explanation results

4.2.1 Global explanations

As shown in Figure 10 and described in Section 3.6, we observe interesting cluster patterns among different coverage types. In this experiment, we utilize basic outlier detection and clustering algorithms to further investigate these patterns.

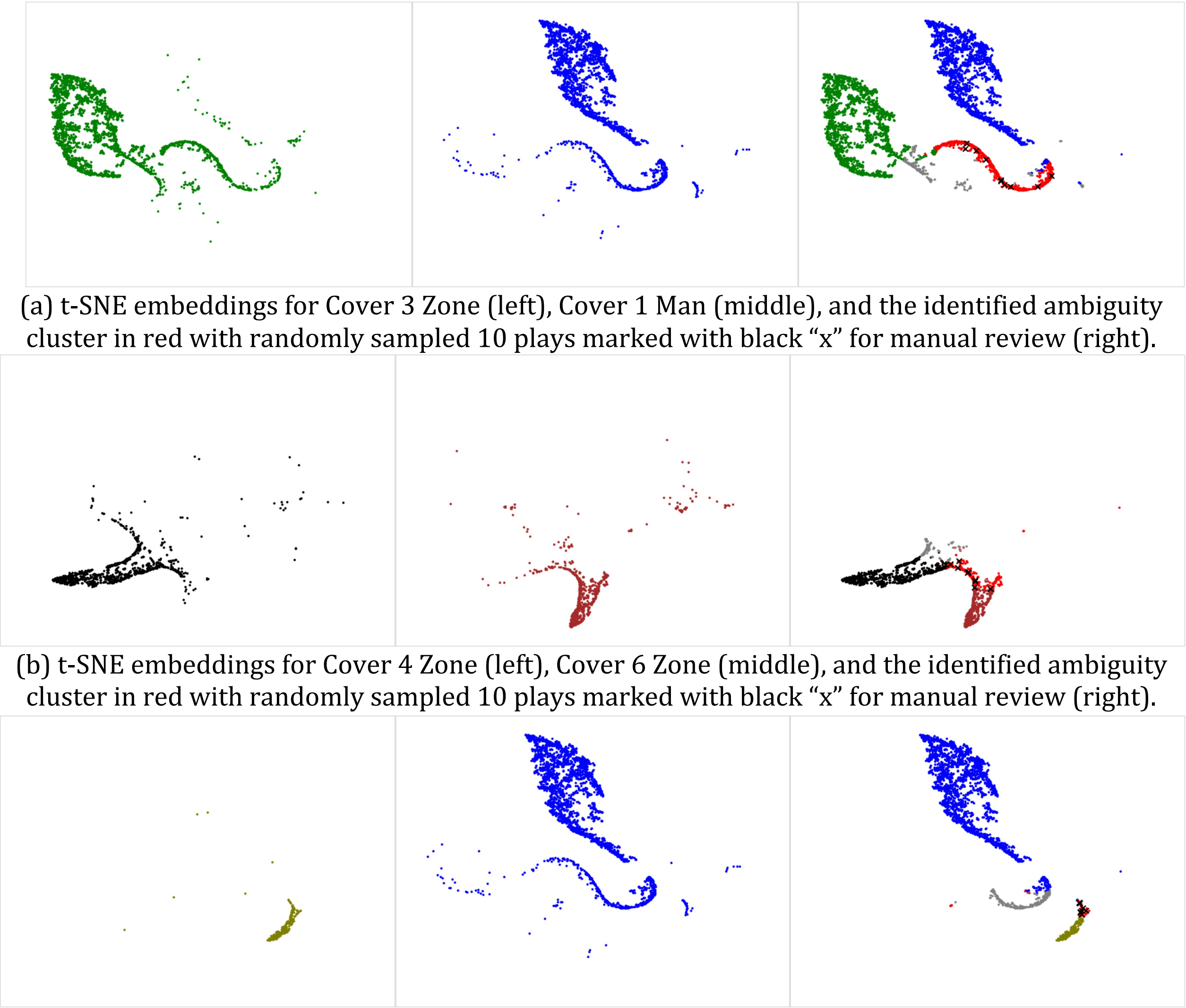
First, we notice that some plays are “mixed” into other coverage types. These plays could potentially be mislabeled and deserve manual inspection. To automatically identify the candidates to review, we design a self-verification method that compares each play’s coverage label with the labels of its neighbors on the learned embedding space. This is achieved with a K-Nearest Neighbors (KNN) classifier. For each example, we compute its correctness score as the classification probabilities on its annotated class label from the KNN. We experimented with different K, *i.e.,* the number of neighbors parameter and chose a relatively large parameter of K=80 to avoid prioritizing samples inside the ambiguity regions. The lowest-score examples are shown in Figure 11. We randomly sampled 13 plays from the highlighted examples for expert review and found that 12 out of the 13 plays were indeed labeled incorrectly. The remaining one play was designated as a zone match split-safety coverage that falls in between Cover 2 Zone (label) and Cover 2 Man (model classification). Inspection of the play footage revealed that the two outside cornerbacks (CBs) kept their eyes on the QB the entire time, which could not be accounted for by the tracking data.

The second interesting observation from Figure 10 is that there are several overlapping regions among the coverage types, indicating coverage ambiguity. We identify the most prominent ambiguities, and utilize a clustering algorithm to extract the associated example plays. Considering the complex topology, we apply spectral clustering algorithm [28] on the play embeddings. We experimented with different number of cluster parameter, by starting with a small value, and gradually increasing it such that the visually identified ambiguity region is covered by one of the clusters. Note that the clustering algorithm is not aimed for the optimal separation of the plays, but rather to effectively select the plays associated with the ambiguity region. The identification results on three prominent regions are visualized in Figure 13. Our expert review uncovered interesting patterns on the adopted coverages:

- The first ambiguity region, as shown in Figure 13(a), deals with the two different single-high coverage concepts: Cover 3 Zone vs Cover 1 Man. The main distinction between these two coverages is man vs zone coverage. Most of the play examples in this region involve some sort of “pattern matching”. In these plays, the coverage responsibilities are contingent upon how the offensive receivers’ routes are distributed, and adjustments can make the play look like a mix of zone and man coverages. For example, one such adjustment we identified applies to Cover 3 Zone, when the cornerback (CB) to one side is locked into man coverage (“Man Everywhere he Goes” or MEG) and the other has a traditional zone drop.
- The second ambiguity region, Cover 4 Zone vs. Cover 6 Zone as shown in Figure 13(b), deals with another pair of coverages that have overlap in their assignments. Cover 6 Zone is best understood as a split field coverage, where one side of the defense is playing Cover 4 Zone and the other is playing Cover 2 Zone. This means one side’s cornerback (the Cover 2 side) is responsible for the “flat” area, while the other is responsible for the deep outside quarter of the field. A key indicator in identifying Cover 6 from Cover 4 will be the flat cornerback initially staying in place or stepping down at snap, while the deep quarter cornerback will start by backpedaling. On a number of plays in this region, the flat area wasn’t threatened by any receivers, so that cornerback eventually had the freedom to drop back, making it look

more like Cover 4. Another pattern from the examples was the relative depth of the safeties. On mostly plays, the defense presents more of a single-high safety shell pre-snap, with one safety significantly deeper than the other. Spacing of the safeties made it appear that they are responsible for a deep half rather than a deep quarter, especially if they are wider.

- A majority of play examples in the third region, Cover 0 Man vs. Cover 1 Man as shown in Figure 13(c), are in the red zone, especially within the 5-yard line. Given the reduced space in this area of the field, it becomes more difficult to determine whether there is a "deep safety" (an indicator of Cover 1 Man). On the plays outside the red zone, the defense showed a single-high safety at snap. However, that player did not drop into the deep middle on any of those plays. Instead, that player would end up in man coverage to replace a blitzing player or help double a dangerous receiver.



(a) t-SNE embeddings for Cover 3 Zone (left), Cover 1 Man (middle), and the identified ambiguity cluster in red with randomly sampled 10 plays marked with black "x" for manual review (right).

(b) t-SNE embeddings for Cover 4 Zone (left), Cover 6 Zone (middle), and the identified ambiguity cluster in red with randomly sampled 10 plays marked with black "x" for manual review (right).

(c) t-SNE embeddings for Cover 0 Man (left), Cover 1 Man (middle), and the identified ambiguity cluster in red with randomly sampled 10 plays marked with black "x" for manual review (right).

Figure 13. Ambiguity analysis on the 3 prominent overlapping regions from t-SNE embeddings: Cover 3 Zone vs. Cover 1 Man (a), Cover 4 Zone vs. Cover 6 Zone (b), and Cover 0 Man vs. Cover 1 Man (c).

4.2.2 Instance explanations

We demonstrate the instance explanation results in this subsection. We first inspect the extracted explanations for “easier” examples whose coverage strategy is clear, to verify that the explanations capture the meaningful player interactions. Then, we utilize the explanation method to shed light on model’s decision on some low-confidence plays.

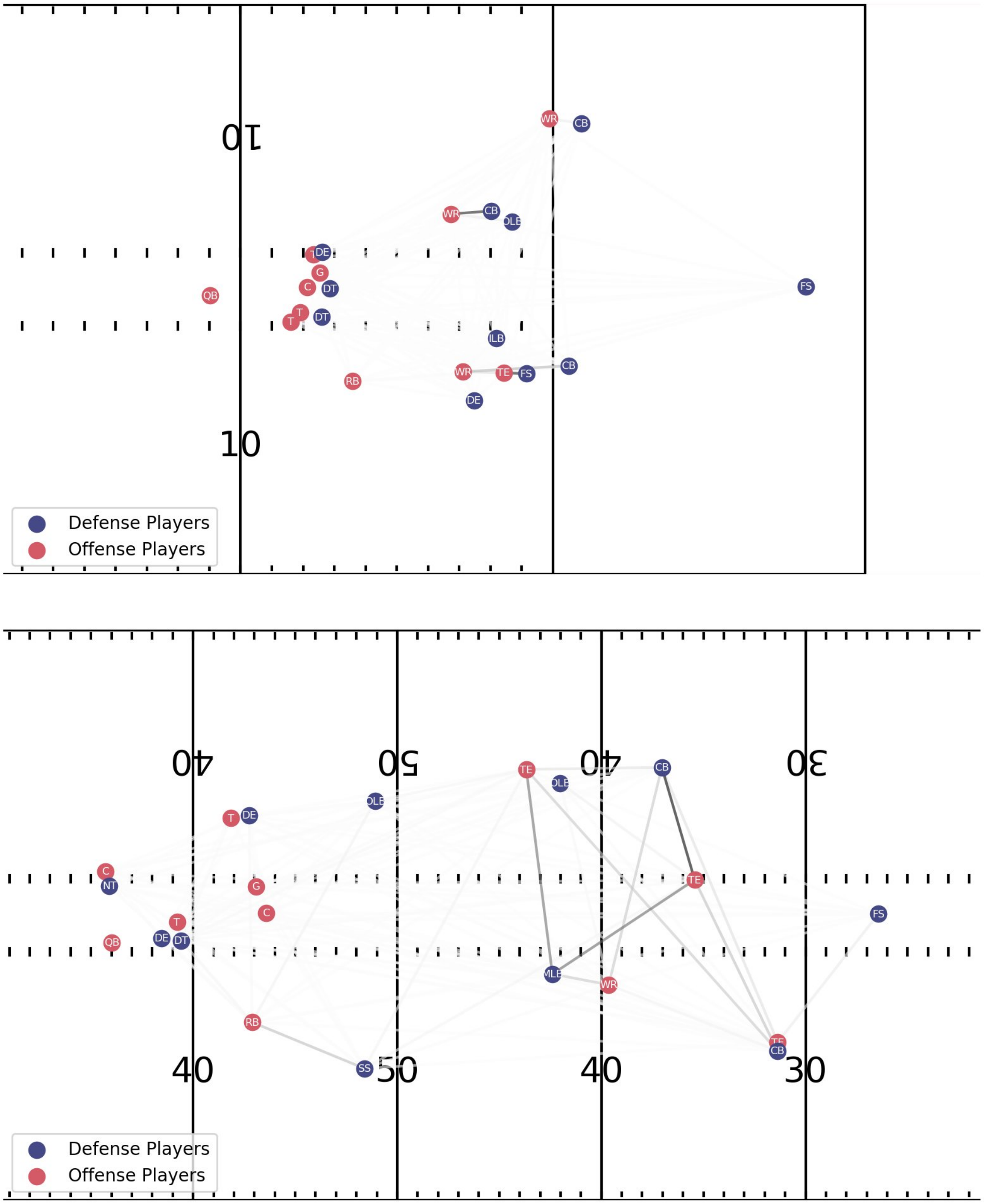


Figure 14. Instance explanation results on a Cover 1 Man play (top) and a Cover 3 Zone play (bottom).

Figure 14 visualizes the instance explanations of a Cover 1 Man play and a Cover 3 Zone play. Note that the frames are selected using the method described in Section 3.6 and Figure 12. On the top figure, the explanation picks up the frame 2.7 seconds into the play and the strong interaction identified by the model between the left slot WR and slot CB. This is aligned with the clear indicator of man coverage with the CB squaring up on the receiver and following him inside and then outside on a whip route. To the other side of the formation, the explanation correctly identifies that the two defensive backs follow the receivers they align across from even as the receivers switch inside and outside, a key man coverage indicator. The TE aligned in the slot is followed by FS on an out-breaking route, while the WR aligned wide is followed by the CB on an in-breaking route. When we consider the deep middle FS, the play is clearly Cover 1 Man.

On the bottom plot of Figure 14, the initial drops of both outside corners to the outside thirds without any regard for the routes being run clearly shows Cover 3 Zone. The explanation picks up this frame at 5.5 seconds into the play, when the pass rush has forced the QB to scramble, but each of these deep third players have maintained their responsibilities. The strong interaction between the TE in the deep middle of the field and both the MLB and CB is the correct reasoning of a Cover 3 framework: he wouldn't be that open if the defense was playing man or match coverage. At the same time, the MLB having strong interactions with both inside TEs who aligned on his side of the formation pre-snap is another clear piece of evidence: he is in zone so he did not follow either TE, even as they entered and exited the area he was responsible for.

After confirming the utility of the instance explanation method, we utilize it to shed light of model's decision when the prediction confidence is low. These plays deserve manual inspection and the instance explanation can help speed up the process. Figure 15 demonstrates the explanation result of a play where the model identifies Cover 1 Man with 61.7% probability and Cover 0 Man with 28.4% probability. When asked to explain the decision of Cover 1 Man, the algorithm identifies the frame (Figure 15 top plot) that comes after the play action fake. At that point it is clearer that the SS is patrolling the deep middle. The highlighted interactions between WR and CB are indeed the correct evidences of man coverage. When asked to explain Cover 0 Man, the algorithm picks up the frame (Figure 15 bottom plot) that comes significantly earlier in the play, right after the snap as the quarterback has turned his back to fake the handoff to the RB. The highlighted interaction between the SS and the right WR is due to the safety moving in that direction, which may have led the model to think he is in man coverage instead of playing the deep middle. This play also conforms to our findings from the third ambiguity region (Figure 13(c)): the condensed space given the proximity to the goal line makes it harder for the model to identify whether there is a "deep safety".

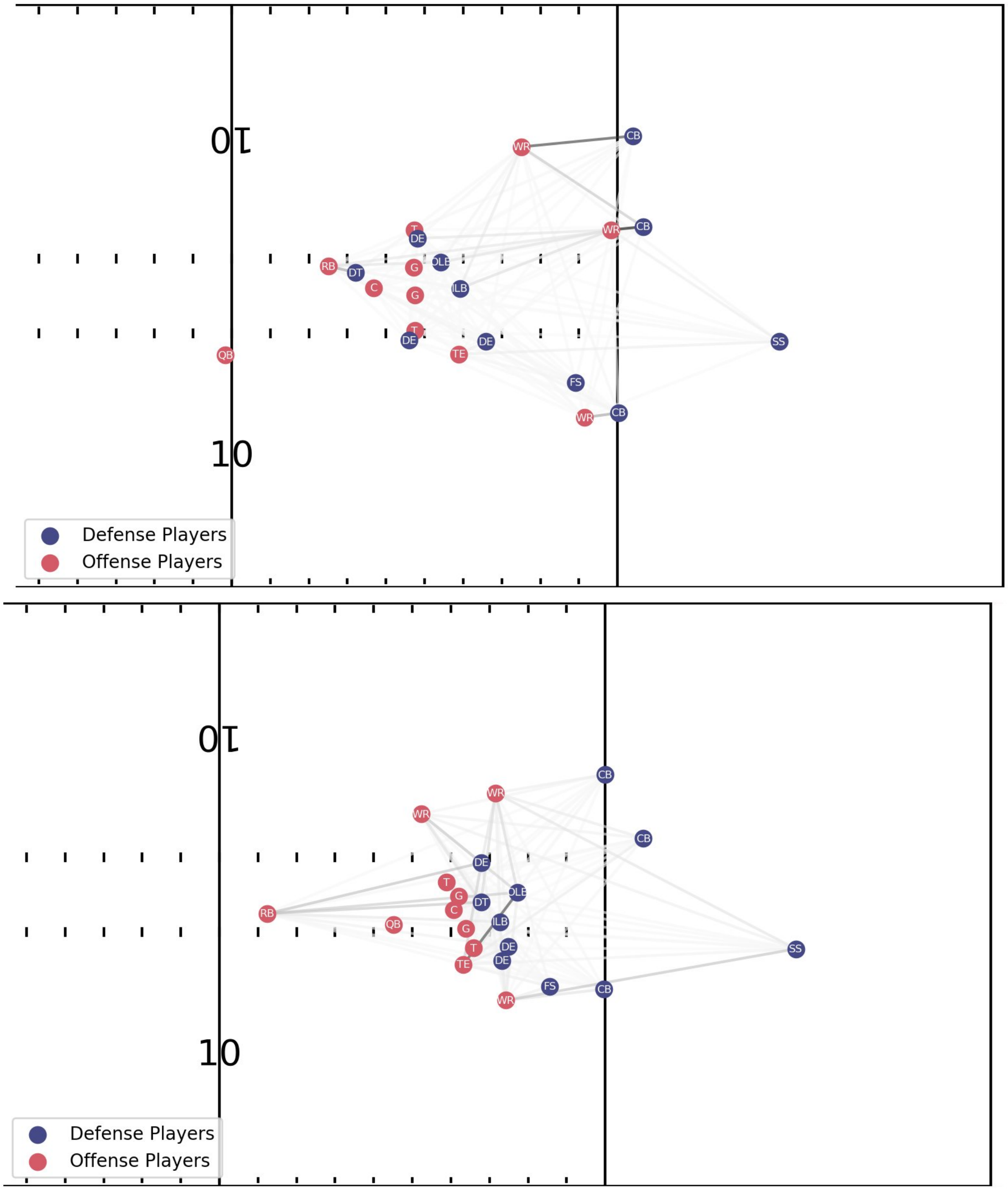


Figure 15. Instance explanation result on a play with 61.7% predicted probability of Cover 1 Man (top) and 28.4% predicted probability of Cover 0 Man (bottom).

Looking back at the play we illustrated in Figure 1, the model predicted Cover 3 Zone with 44.5% probability and Cover 1 Man with 31.3% probability. We generate the explanation results for both classes as shown in Figure 16. The top plot for Cover 3 explanation comes right after the ball snap. The CB on the offense's right has the strongest interaction lines, because he is facing the QB and stays in place. He ends up squaring off and matching with the receiver on his side who threatens him deep.

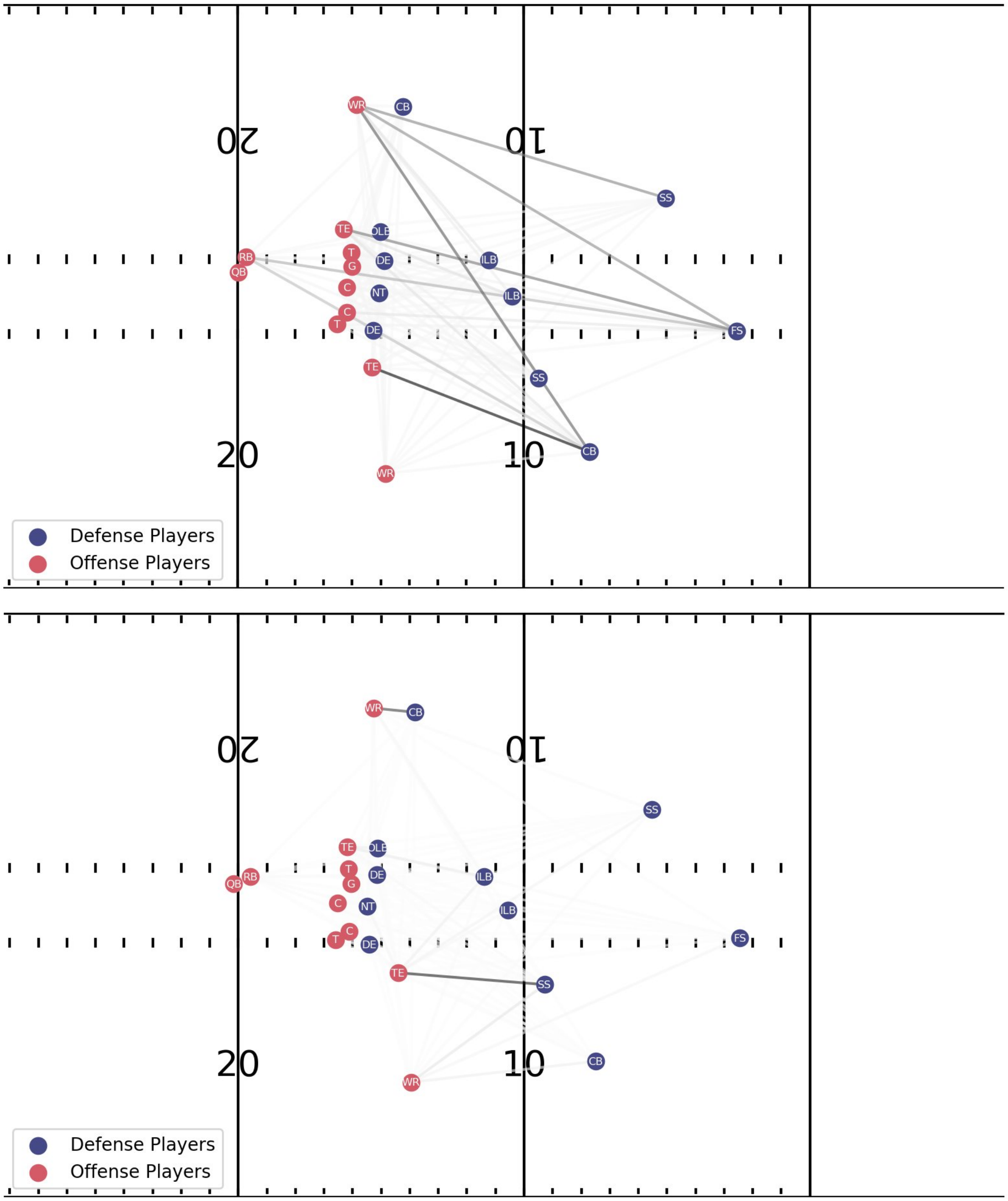


Figure 16 Instance explanation result on a play with 44.5% predicted probability of Cover 3 Zone (top) and 31.3% predicted probability of Cover 1 Man (bottom). This is the same play as the one we illustrated in Figure 1.

The bottom plot for Cover 1 explanation comes a moment later, as the play action fake is happening. One of the strongest interactions is with the CB to the offense's left, who is dropping with the WR. Play footage reveals that he keeps his eyes on the QB before flipping around and running with the WR who is threatening him deep. The SS on the offense's right also has a strong interaction with the TE on his side, as he starts to shuffle as the TE breaks inside. He ends up following him across the formation, but the TE starts to block him, indicating the play was likely a run-pass option. This explains the uncertainty of the model's classification: the TE is sticking with the SS by design, creating biases in the data.

5. Conclusion

This paper presents a novel ensemble CNN-attention model to classify defense coverage schemes in a performant manner. It significantly outperformed existing frame-based model and achieved production-ready performance. This approach is easily generalizable and extensible to include additional types of coverages beyond the eight coverages we considered in the paper. The classification model has been deployed to production by NFL NGS engineering and product teams.

To extract insights regarding coverage ambiguity and model decision-making process, we further developed a comprehensive model explanation method. Through global explanation that uncovers coverage ambiguity patterns and instance explanation that highlights critical signals on the player interactions, our approach revealed interesting insights about the team and player behaviors. This also enables intelligent selection of plays for efficient human reviews.

In future work, we plan to investigate game-theoretic approaches [23, 24] for the explanation of the coverage classification model. In addition, we would like to study temporal graph neural networks (GNNs) that can directly model the player interactions from raw data, as well as GNN-based model explanation approaches [25].

6. Acknowledgements

We would like to thank Keegan Abdoo from NFL NGS for providing the invaluable football insights, and Huzefa Rangwala from Amazon ML Solutions Lab for paper writing suggestions.

References

[1] Eric Eager, George Chahrouri, Timo Riske, Brad Spielberger, LauSze Yui, Zach Drapkin, Tej Seth. “Using Tracking and Charting Data to Better Evaluate NFL Players: A Review” In MIT Sloan Sport Analytics Conference. 2022.

[2] Horton, Michael. "Learning feature representations from football tracking." In MIT Sloan Sports Analytics Conference. 2020.

[3] Hochstedler, Jeremy H. "Incorporating spatiotemporal machine learning into Major League Baseball and the National Football League." PhD diss., Massachusetts Institute of Technology, 2016.

[4] Goyal, Udgam. "Leveraging machine learning to predict playcalling tendencies in the NFL." PhD diss., Massachusetts Institute of Technology, 2020.

[5] Reyers, Matthew, and Tim B. Swartz. "Quarterback evaluation in the national football league using tracking data." AStA Advances in Statistical Analysis (2021): 1-16.

[6] da Silva, Gustavo Pompeu, and Rafael de Andrade Moral. "Frame by frame completion probability of an NFL pass." arXiv preprint arXiv:2109.08051 (2021).

[7] Lin Lee Cheong, Xiangyu Zeng and Ankit Tyagi. “Prediction of Defensive Player Trajectories in NFL Games with Defender CNN-LSTM Model” In MIT Sloan Sports Analytics Conference. 2021.

[8] Ben Baldwin. “Computer Vision with NFL Player Tracking Data using torch for R: Coverage classification Using CNNs.” https://www.opensourcefootball.com/posts/2021-05-31-computer-vision-in-r-using-torch/

[9] Skoki, Arian, Jonatan Lerga, and Ivan Štajduhar. "ML-Based Approach for NFL Defensive Pass Interference Prediction Using GPS Tracking Data." In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 1038-1043. IEEE, 2021.

[10] Raabe, Dominik, Reinhard Nabben, and Daniel Memmert. "Graph representations for the analysis of multi-agent spatiotemporal sports data." Applied Intelligence (2022): 1-21.

[11] Dutta, Rishav, Ronald Yurko, and Samuel L. Ventura. "Unsupervised methods for identifying pass coverage among defensive backs with NFL player tracking data." Journal of Quantitative Analysis in Sports 16, no. 2 (2020): 143-161.

[12] Dickmanns, Ludwig. "Pose Estimation and Analysis for American Football Videos." (2021).

[13] Joash Fernandes, Craig, Ronen Yakubov, Yuze Li, Amrit Kumar Prasad, and Timothy CY Chan. "Predicting plays in the national football league." Journal of Sports Analytics 6, no. 1 (2020): 35-43.

[14] Lalwani, Abhinav, Aman Saraiya, Apoorv Singh, Aditya Jain, and Tirtharaj Dash. "Machine Learning in Sports: A Case Study on Using Explainable Models for Predicting Outcomes of Volleyball Matches." arXiv preprint arXiv:2206.09258 (2022).

[15] Silver, Joshua, and Tate Huffman. "Baseball Predictions and Strategies Using Explainable AI." In The 15th Annual MIT Sloan Sports Analytics Conference. 2021.

[16] Wang, Yuanchen, Weibo Liu, and Xiaohui Liu. "Explainable AI techniques with application to NBA gameplay prediction." Neurocomputing 483 (2022): 59-71.

[17] Dmitry Gordeev, Philipp Singer. “1st place solution The Zoo.” https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/119400

[18] Tej Seth, Ryan Weisman, “PFF Data Study: Coverage scheme uniqueness for each team and what that means for coaching changes”, https://www.pff.com/news/nfl-pff-data-study-coverage-scheme-uniqueness-for-each-team-and-what-that-means-for-coaching-changes

[19] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based

localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.

[20] NFL Football Operations, "Which NFL teams mix up defensive coverages the most week-to-week", https://operations.nfl.com/gameday/analytics/stats-articles/which-nfl-teams-mix-up-defensive-coverages-the-most-week-to-week/

[21] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[22] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.

[23] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[24] Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees." Nature machine intelligence 2, no. 1 (2020): 56-67.

[25] Ying, Zhitao, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. "Gnnexplainer: Generating explanations for graph neural networks." Advances in neural information processing systems 32 (2019).

[26] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. 11 (2008).

[27] Springenberg, Jost Tobias, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. "Striving for simplicity: The all convolutional net." arXiv preprint arXiv:1412.6806 (2014).

[28] Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17, no. 4 (2007): 395-416.

Appendix

Player position acronyms in Figure 2

Defensive positions

W	"Will" Linebacker, or the **w**eak side LB
M	"Mike" Linebacker, or the **m**iddle LB
S	"Sam" Linebacker, or the **s**trong side LB
CB	Cornerback
DE	Defensive End
DT	Defensive Tackle
NT	Nose Tackle
FS	Free Safety
SS	Strong Safety
S	Safety

Offensive positions

X	Usually the number 1 WR in an offense, they align on the LOS. In trips formations, this receiver will often align isolated on the backside.
Y	Usually the starting TE, this player will often align in-line and to the opposite side as the X.
Z	Usually more of a slot receiver, this player will often align off the LOS and on the same side of the field as the TE.
H	Traditionally a fullback, this player is more often a third WR or a second TE in the modern league. They can align all over the formation, but are almost always off the line of scrimmage. Depending on the team, this player could also be designated as a F.
T	The featured running back. Other than empty formations, this player will align in the backfield and be a threat to receive the handoff.
QB	Quarterback
C	Center
G	Guard

Player position acronyms in other figures, if not in the above

Defensive positions

LB	Linebacker
ILB	Inside Linebacker
OLB	Outside Linebacker
MLB	Middle Linebacker

Offensive positions

RB	Running Back
FB	Fullback
WR	Wide Receiver
TE	Tight End
LG	Left Guard
RG	Right Guard
T	Tackle
LT	Left Tackle
RT	Right Tackle