Ahmedabad University

CSE623 : Machine Learning | The Overfitters

Identifying hard stop & momentary stop using vehicle trajectory dataset

Vishesh Bhatia
AU2240027

Jay Raval
AU2240151

Jinil Savaj
AU2240159

Karan Prajapati
AU2240161

Meet Suthar
AU2240198

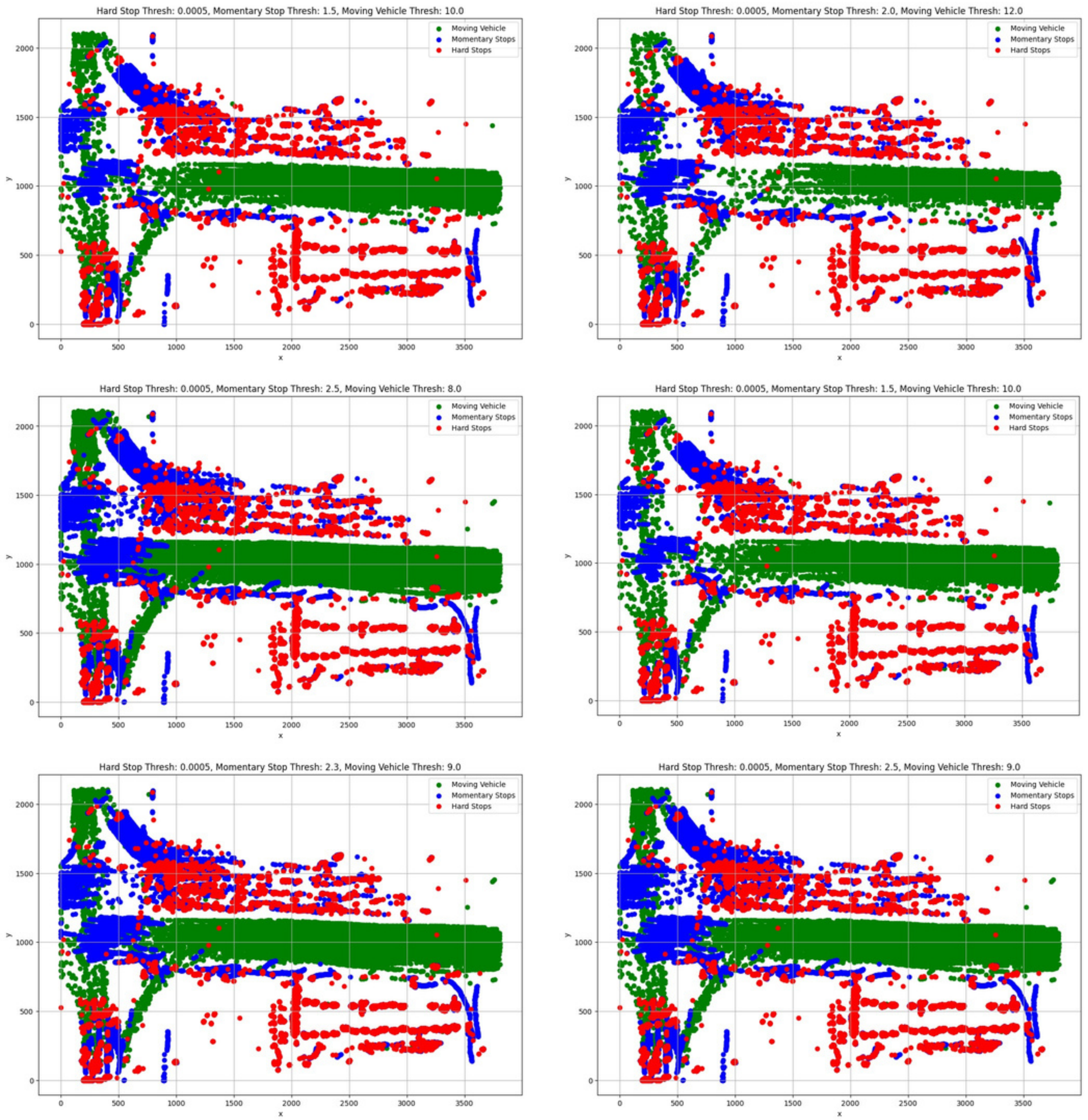
Problem Statement

- Given a dataset containing extracted vehicle trajectories at a signalized intersection to analyze different stopping behaviors.
- At intersections, we assume drivers follow traffic signals, leading to vehicles stopping and restarting at red lights, which we classify as a momentary stop.
- Some vehicles remain stationary for an extended period, such as parked vehicles, representing a hard stop.
- There are vehicles that continue moving without stopping, categorized as moving vehicles.
- Our objective is to develop a machine learning model to accurately distinguish between these three conditions using trajectory data.

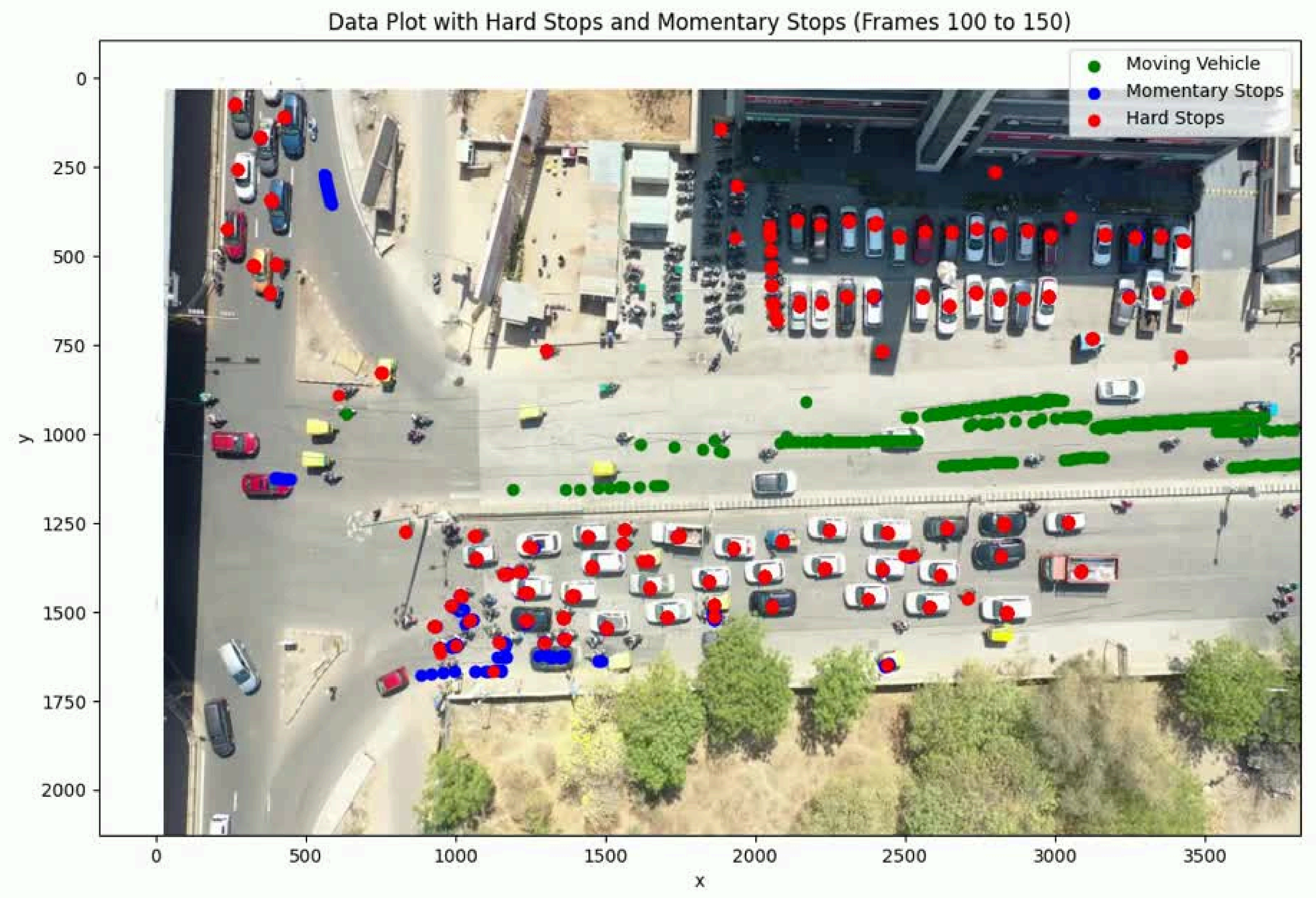
Instructor Feedback

```
hard_stop_thresholds = [0.0005, 0.0007, 0.0009, 0.001, 0.0011, 0.0013, 0.0015]
momentary_stop_thresholds = [1.5, 1.7, 1.9, 2.0, 2.1, 2.3, 2.5]
moving_vehicle_thresholds = [8.0, 9.0, 10.0, 11.0, 12.0]
```

- Till Midsem, we were labelling the given dataset by taking different thresholds, plot all the combinations and chose appropriate threshold values by eyeballing graphs.
- We, were told to try different labelling methods and compare results with DBSCAN model.
- Implementing a bounding-box to see the change in cluster size of momentary stops with time when signal starts and stops.



MidSem Results

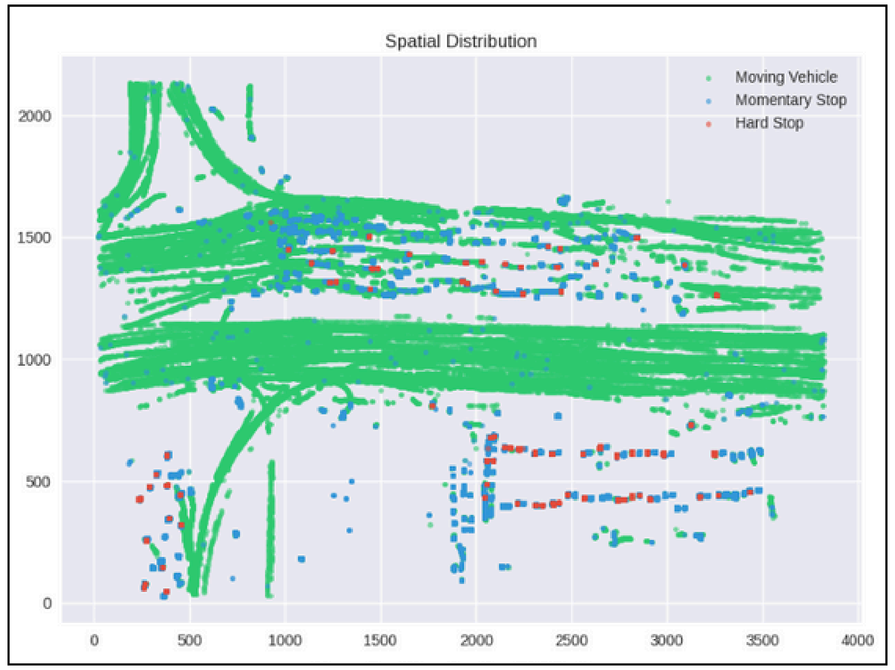


Accuracy Score: 0.8529680365296803
Classification Report:

	precision	recall	f1-score	support
0	0.89	0.92	0.91	87718
1	0.84	0.87	0.85	14991
2	0.52	0.39	0.44	13361
accuracy			0.85	116070
macro avg	0.75	0.73	0.73	116070
weighted avg	0.84	0.85	0.85	116070

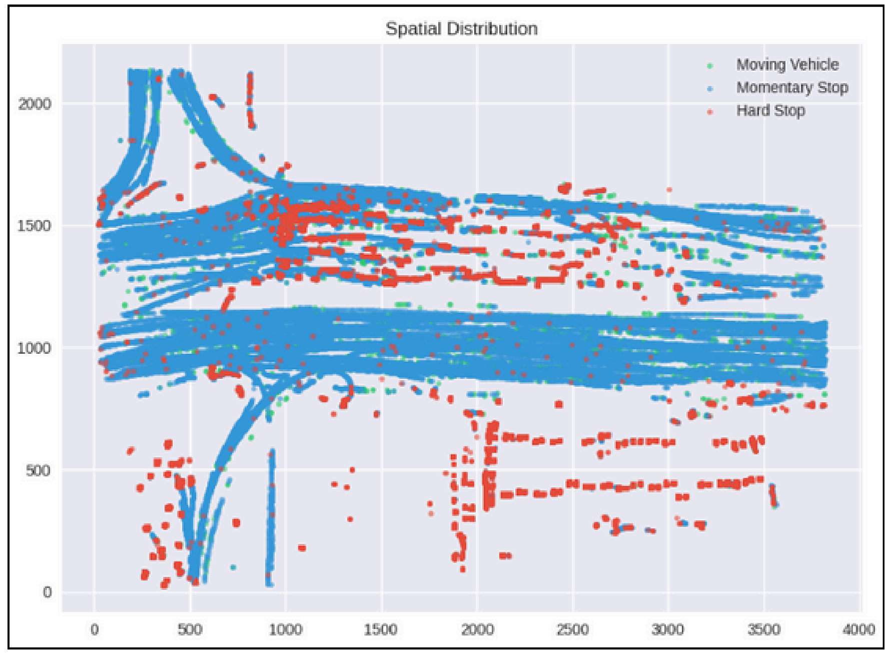
Approach

Different Labeling Method Used



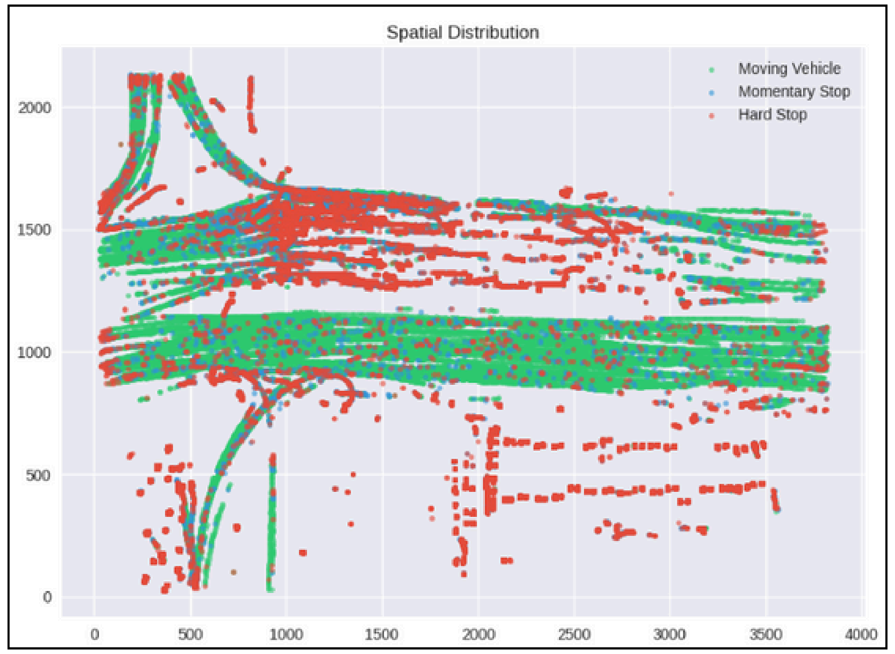
Time-Based

Improved stop detection using frame durations and event grouping (Used in final model)



HMM

Sequential modeling with KMeans-initialized Gaussian HMM.



Clustering

KMeans with speed, acceleration, rolling average, outlier handling

Changepoint

Segments based on speed signal using Ruptures (PELT)

Results

MidSem Approach Improvement

```python
    elif model_type == 'xgb':
        steps.extend([
            ('smote', SMOTE(random_state=42, sampling_strategy=sampling_strategy)),
            ('classifier', XGBClassifier(
                n_estimators=400,
                max_depth=6,
                learning_rate=0.1,
                subsample=0.8,
                colsample_bytree=0.8,
                objective='multi:softmax',
                num_class=3,
                scale_pos_weight=[1, class_counts[0]/class_counts[1], class_counts[0]/class_counts[2]],
                random_state=42,
                n_jobs=-1
            ))
        ])

    return Pipeline(steps)
```

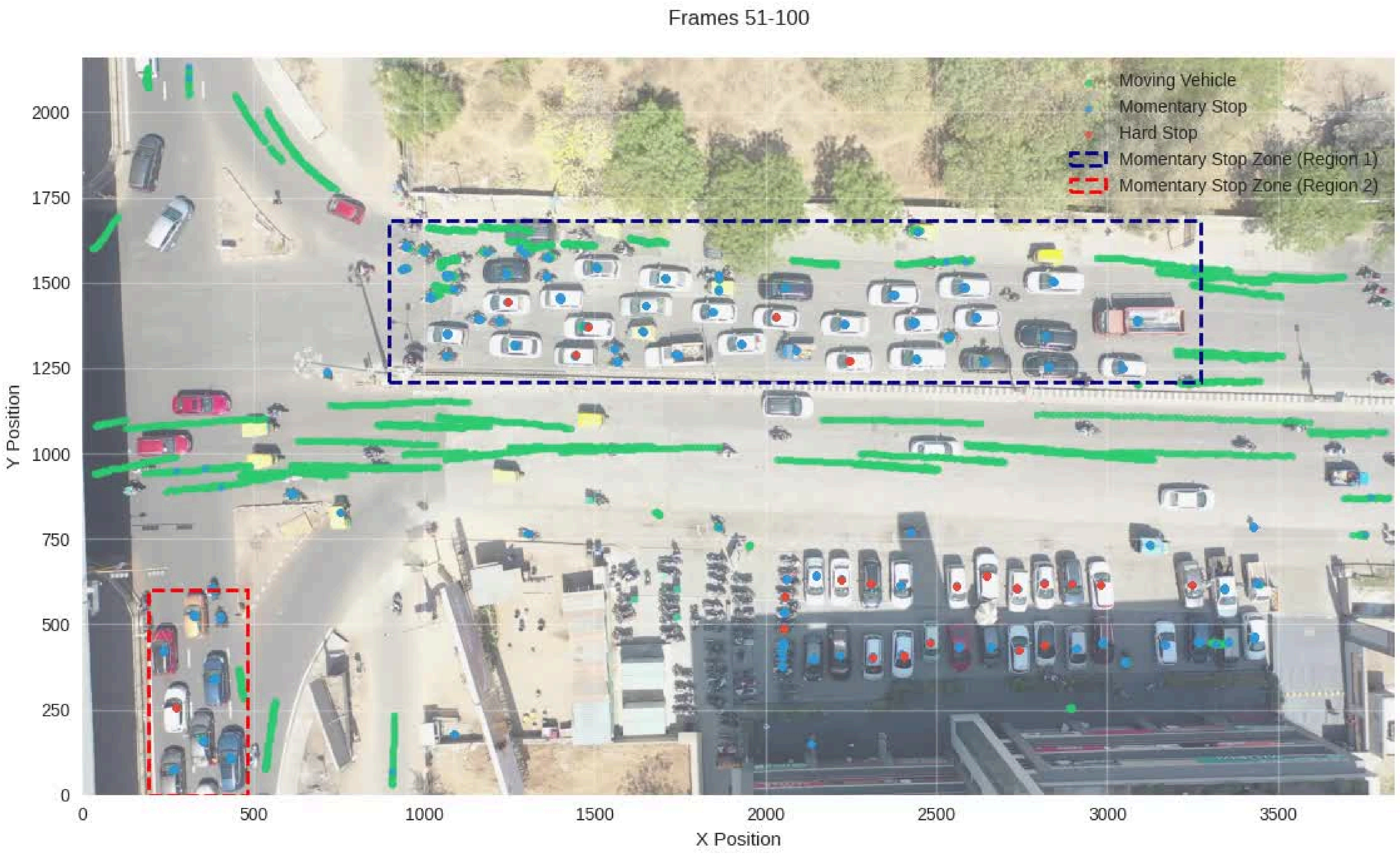
```
Accuracy: 0.8783
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.85      0.92     16319
           1       0.70      0.97      0.81      2619
           2       0.62      0.95      0.75      2353

    accuracy                           0.88     21291
   macro avg       0.77      0.92      0.83     21291
weighted avg       0.91      0.88      0.89     21291
```

Results

Video with Dynamic Bounding Box




```
Model Performance:
Accuracy: 0.81

Classification Report:
                precision    recall  f1-score   support

Moving Vehicle       0.85      0.85      0.85     18053
Momentary Stop       0.76      0.77      0.77     12840
     Hard Stop       0.67      0.65      0.66      1448

      accuracy                           0.81     32341
     macro avg       0.76      0.76      0.76     32341
  weighted avg       0.81      0.81      0.81     32341


Confusion Matrix:
[[15335  2647    71]
 [ 2551  9892   397]
 [   65   443   940]]
```

Endsem Model Performance

Future Work

- **Model Generalization Across Geographies:** Evaluate the model on datasets from different cities or road types to check its adaptability across varied traffic conditions, driving behaviors, and road infrastructures.

- **Multi-Model Ensemble Approach:** Combine the strengths of multiple models (e.g., XGBoost, Random Forest, SVM) through ensembling to improve robustness and reduce model bias.

- **Data Augmentation and Synthetic Data Generation:** Generate synthetic driving patterns or stops using GANs or simulation tools to augment the training data, especially for rare events.

References

[1] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: https://doi.org/10.1023/a:1010933404324

[2] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” Applied Statistics, vol. 28, no. 1, p. 100, 1979, doi: https://doi.org/10.2307/2346830

[3] L. E. Baum and T. Petrie, “Statistical Inference for Probabilistic Functions of Finite State Markov Chains,” The Annals of Mathe matical Statistics, vol. 37, no. 6, pp. 1554–1563, Dec. 1966, doi: https://doi.org/10.1214/aoms/1177699147

[4] G. Y. Oukawa, P. Krecl, and A. C. Targino, “Fine-scale modeling of the urban heat island: A comparison of multiple linear regression and random forest approaches,” Science of The Total Environment, vol. 815, p. 152836, Apr. 2022, doi: https://doi.org/10.1016/j.scitotenv.2021.152836