DATA 606 - Capstone Project

Data Fusion for Predicting Highway Maintenance and Deterioration Trends

Team Members:
Chandra Sekhar Katipalli
Sindura Reddy Challa
Sanjana Reddy Soma

Problem Statement

Road infrastructure degrades over time—especially under the strain of freight traffic—resulting in **costly repairs, safety risks, and travel disruptions**.

Yet, most maintenance strategies today are **reactive**, addressing damage *after* it happens.

Our goal: **Predict deterioration before it occurs** — enabling **proactive maintenance**, **reduced failures** and **smarter resource planning**.

Why We Care

Poor Road Conditions Have Real Costs

U.S. drivers lose $1,400+ per year on average due to poor road conditions (vehicle damage, fuel, delays)

Rough roads contribute to 1 in 3 traffic fatalities in some states

Aging infrastructure → rising maintenance costs & safety concerns

Current Practices Are Reactive

Most maintenance decisions are based on visual inspections and annual reports

Lack of early prediction leads to unplanned failures and higher repair costs

UMBC

Research Questions

1. Does combining HPMS and FAF data improve predictions of road roughness (IRI) compared to using just one dataset?
2. Which features are most important in predicting road conditions — like traffic volume, freight load, or pavement quality?
3. Can our model reduce forecasting errors and help plan maintenance more efficiently?
4. What's the best way to predict IRI at different levels — for full routes vs. smaller road segments?
5. How should we present our findings to make them useful for transportation planners and highway maintenance teams?

Data Collection & Sources

Freight Analysis Framework (FAF4.5) Dataset
• **Source:** U.S. Bureau of Transportation Statistics (FAF4.5) • **Data Format:** Shapefiles • **Total Variables:** 16 • **Years Used:** 2013 – 2022 • **Purpose:** Captures annual freight movement (tons, value, miles, etc.)

Data Collection & Sources

Highway Performance Monitoring System (HPMS) Dataset
• **Source:** U.S. Federal Highway Administration (HPMS Data) • **Data Format:** Shapefiles • **Total Variables:** Varies by state submissions • **Years Used:** 2013 – 2022 • **Purpose:** Provides detailed roadway condition and usage data

FAF Dataset: Tracking Freight Movement

Column	Description	Why it matters
dms_orig	Origin FAF region(where freight movement begins)	Starting point of freight(will be used to link datasets)
dms_dest	Destination FAF region(where freight movement ends)	Ending point of freight
dms_mode	Mode of Transport(Truck, Rail, Air, Water etc.)	Helps determine if it is mode of freight movement roadways, railways or airways
curval, tmiles, tons, value	Freight metrics (current value, ton-miles, volume, and monetary value)	

HPMS Dataset: Monitoring Pavement Conditions

Column	Description	Why it matters
IRI_VN	International Roughness Index	Helps determine the condition of the road.
AADT_VN	Annual Average Daily Traffic	Average daily traffic movement on the section of road.
IS_IMPROVED	Flag indicating if the segment was improved since the last year	
THROUGH_LA	Number of through lanes	Shorter lanes can experience more deterioration.
SPEED_LIMI	Speed limit on the road segment	

What is IRI?

IRI is the road roughness index most commonly used worldwide for evaluating and managing road systems. Road roughness is the primary indicator of the utility of a highway network to road users. IRI is defined as a statistic used to estimate the amount of roughness in a measured longitudinal profile. IRI units of either m/km or in/mi.

IRI Threshold for Maintenance

- IRI > 94 inches/mile typically indicates corrective action is needed.

Data Preprocessing

Freight Analysis Framework Data

- Filtered by mode of transport (Truck).
- Considered only in-state movements.
- Few key metrics: tons, value, tmiles, curval
- Applied tons-weighted averages for each year (2013–2022)
- Merged yearly files into a state-level aggregated dataset.

Data Preprocessing

Highway Performance Monitoring System Data

- Considered only interstate roads present in CFS Zone Areas for FAF compatibility. We used county codes as reference.
- Standardized column names and column count across years.
- Handled missing values for column Speed Limit.
- Made sure road sections are consistent across years.
- Merged all years (2013–2022) into a single cleaned HPMS dataset and integrated FAF data.

Feature Engineering

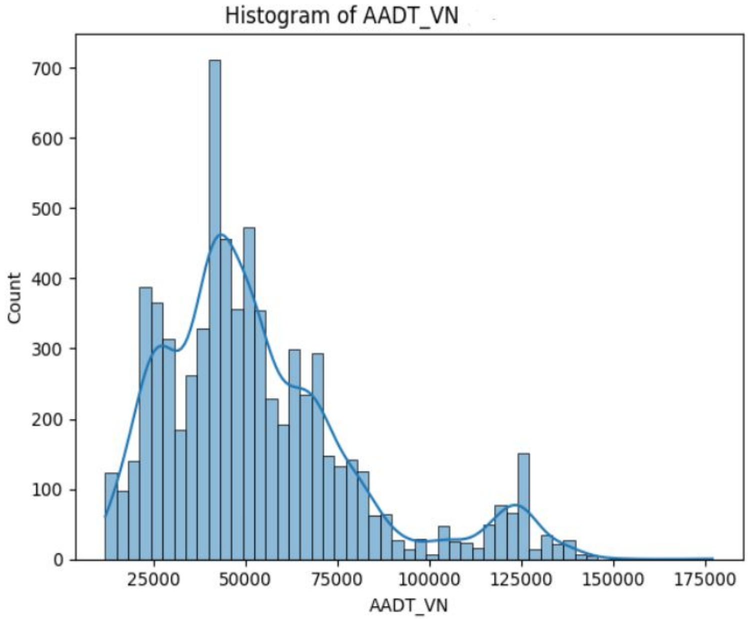
- Rounded BEGIN_POIN and END_POINT to nearest tenth for data consistency across years.
- Created IS_IMPROVED column based on Year_last_improvement and, also if we have seen improvement in IRI value in previous year.
- Did encoding for categorical values such as ROUTE_ID.

Merged HPMS + FAF datasets based on YEAR and COUNTY CODES

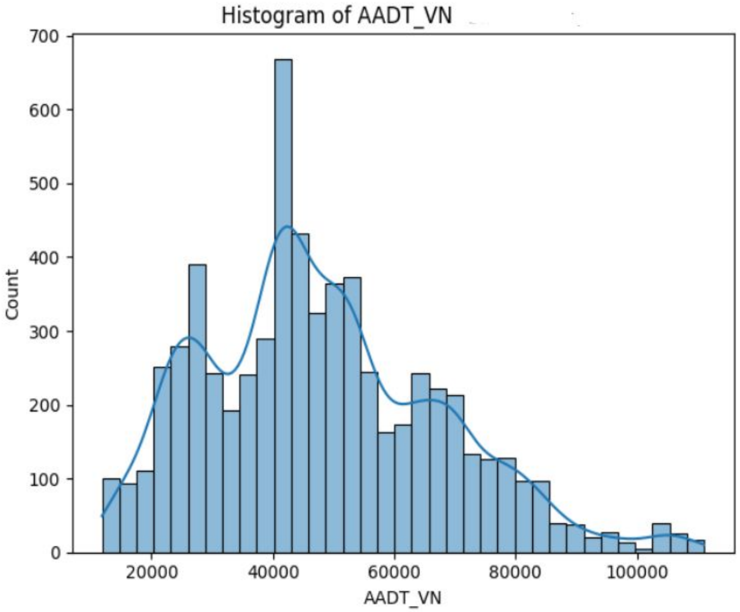
- Aggregated and aligned freight data to corresponding highway segments by county and year

OUTLIERS

Before



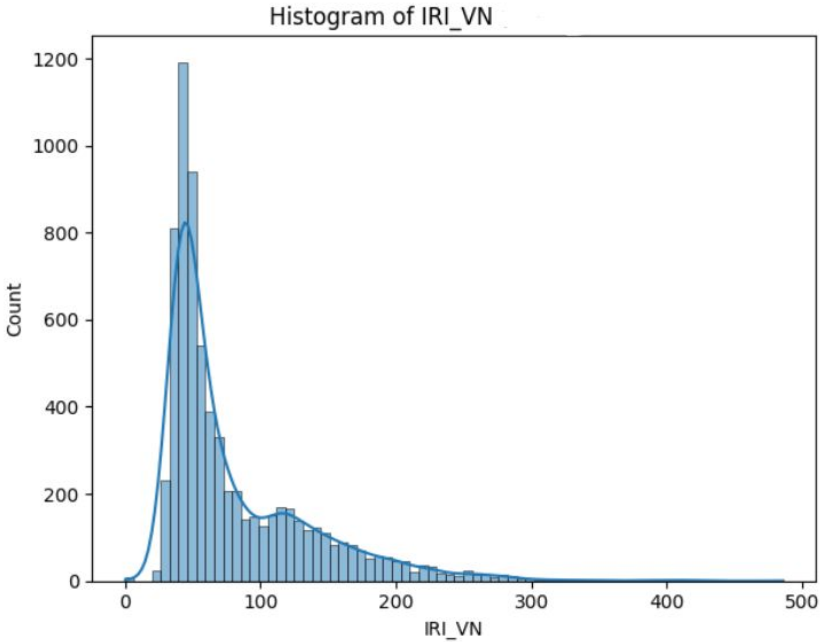
After



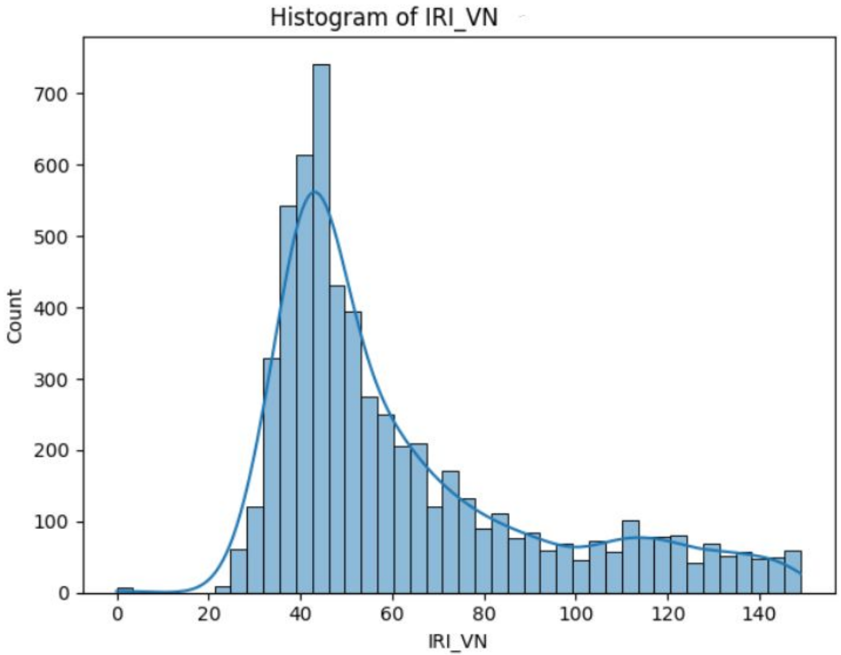
We used Interquartile Range method to remove the outliers. Total outliers count 391 data points.

OUTLIERS

Before

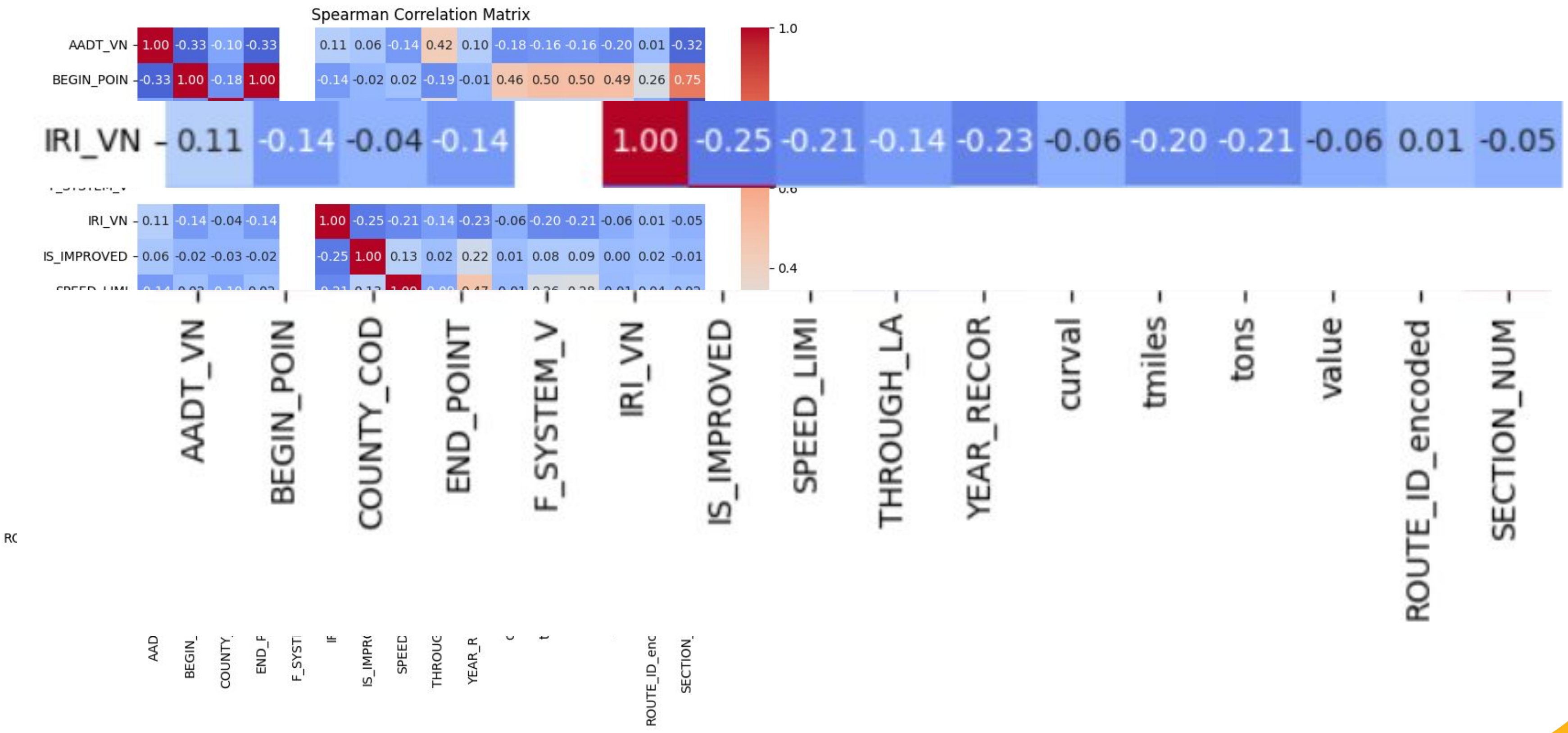


After



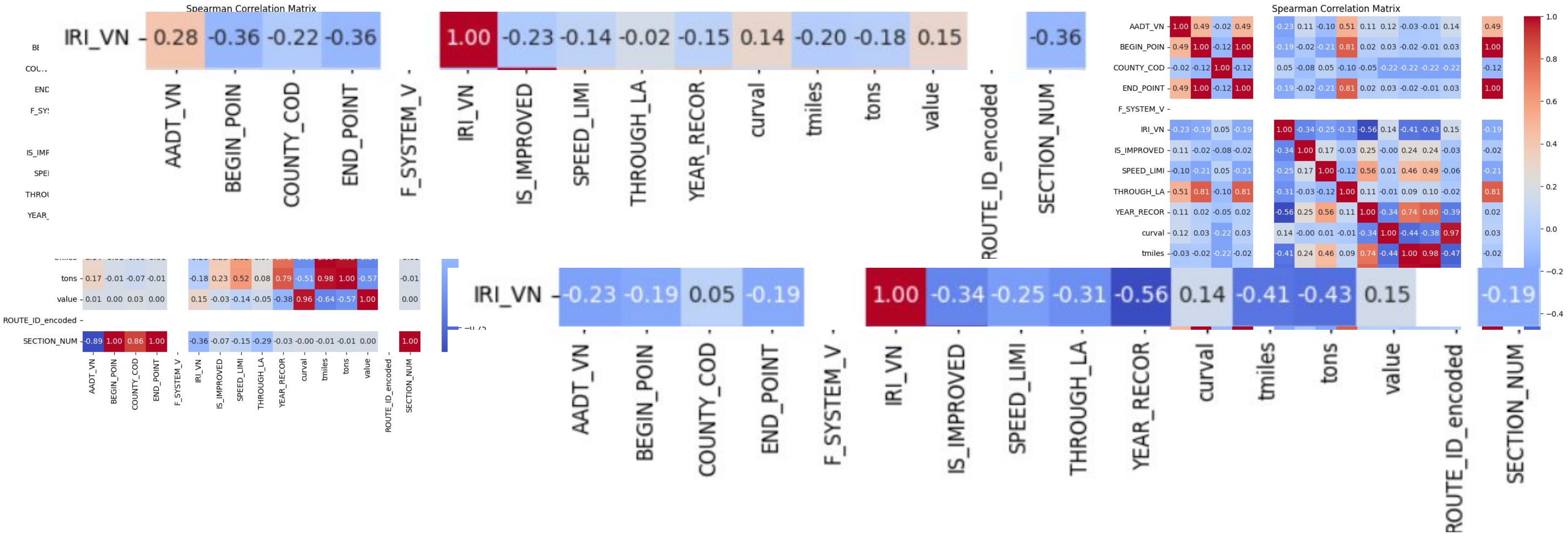
We used Interquartile Range method to remove the outliers. Total outliers count 424 data points.

Exploratory Data Analysis (Whole Dataset)



Exploratory Data Analysis (Route Level)

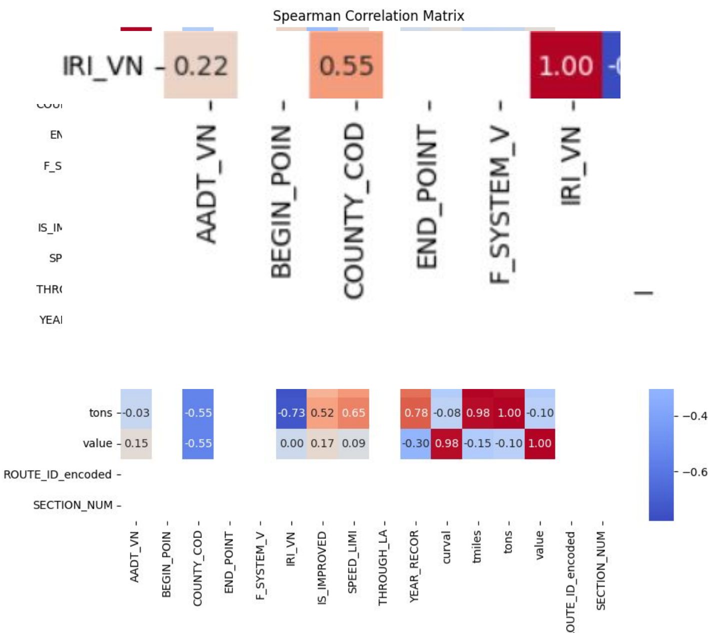
IN0000590000 **IN0004590000**



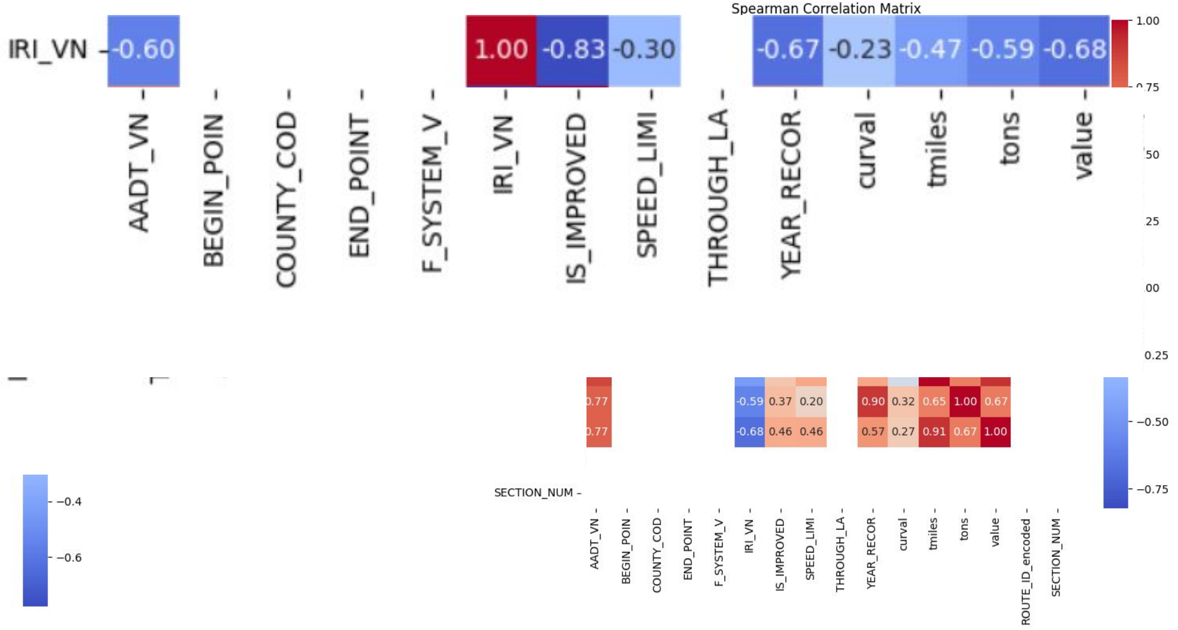
IN0004590000

Exploratory Data Analysis (Section Level)

IN0004590000 (section 1.6 - 1.7)



IN0000100000 (section 38.3 - 38.4)



Model Assessment - After Hyperparameter tuning (Whole Dataset)

Model	MSE	R^2
Linear Regression	652.63	0.138
Decision Tree	537.88	0.289
KNN	461.50	0.390
Random Forest	437.03	0.423
Gradient Boosting	338.69	0.553
Voting Regressor	373.44	0.507
XGBoost	330.27	0.562

- Common models showed high error variance and low accuracy.
- Local patterns were lost when training on combined data.
- Route-wise modeling may capture IRI behaviour.
- So we tried Route-wise modeling.

Route-wise Modeling

Gradient Boosting Parameters after tuning with RandomSearchCV

learning_rate=0.03
n_estimators=600
max_depth=3
subsample=0.85
min_samples_split=4
min_samples_leaf=2
max_features='sqrt'
random_state=42

Model Assessment - After Hyperparameter tuning (Route-Wise)

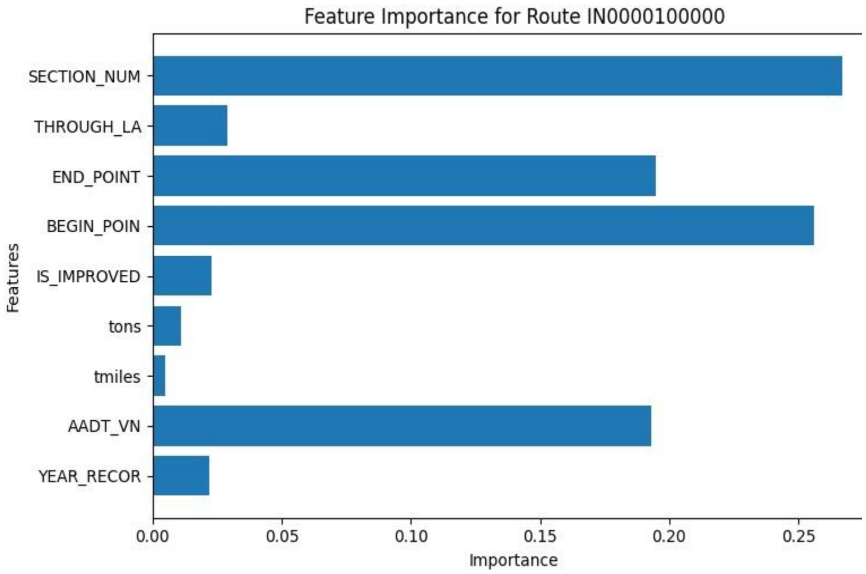
Trained and tuned models **per RouteID**

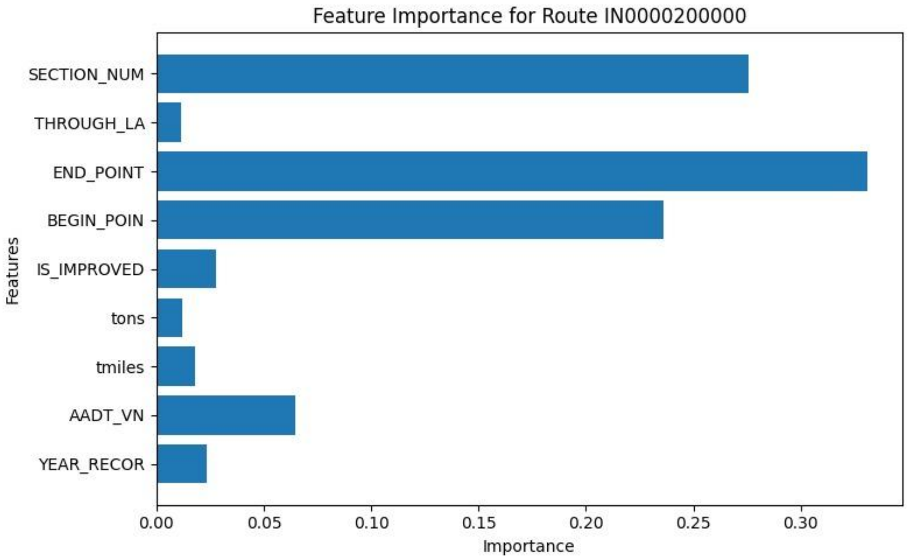
ROUTE_ID	Model	Train MSE	Test MSE	Train R2	Test R2	Train Size	Test Size
IN0000100000	Gradient Boosting	95.31055419	155.2640871	0.914073598	0.861504585	980	285
IN0000100000	Voting Regressor	59.96074957	191.2753859	0.945942907	0.829382541	980	285
IN0000200000	Gradient Boosting	65.16132541	174.7990556	0.894610381	0.744837085	523	153
IN0000200000	Voting Regressor	44.69868763	177.3464134	0.927705926	0.74111858	523	153
IN0000220000	Voting Regressor	77.00933733	282.2153559	0.875910208	0.489643953	401	121
IN0000220000	Gradient Boosting	116.1610751	324.2314621	0.812822651	0.413662355	401	121
IN0000590000	Voting Regressor	83.80391786	481.7638142	0.92636408	0.674882144	789	214
IN0000590000	Gradient Boosting	133.2645836	513.4753012	0.882904518	0.653481677	789	214

Even after **route-wise modeling**, some routes showed **performance limitations.**
Deterioration is a **temporal process** → current condition depends on previous years.
Since one single model is not equalling performing well for all ROUTEs, we decided to convert our data into sequential data and train neural networks.

Feature Importance

Gradient Boosting Route-wise Feature Importance





Sequential Modeling

To capture year-over-year trends in IRI and make time-based predictions.

Conversion to Sequential Format:

Window Size: 8 years

Target: IRI of the following year

Format:

Input: Past 8 years of features

Output: IRI of year t+1

Base Layer

<u>Before tuning:</u>

Dense layers: 3

Dropout : 0.1, 0.3

Activation: relu

Optimizer: adam

Loss: mse

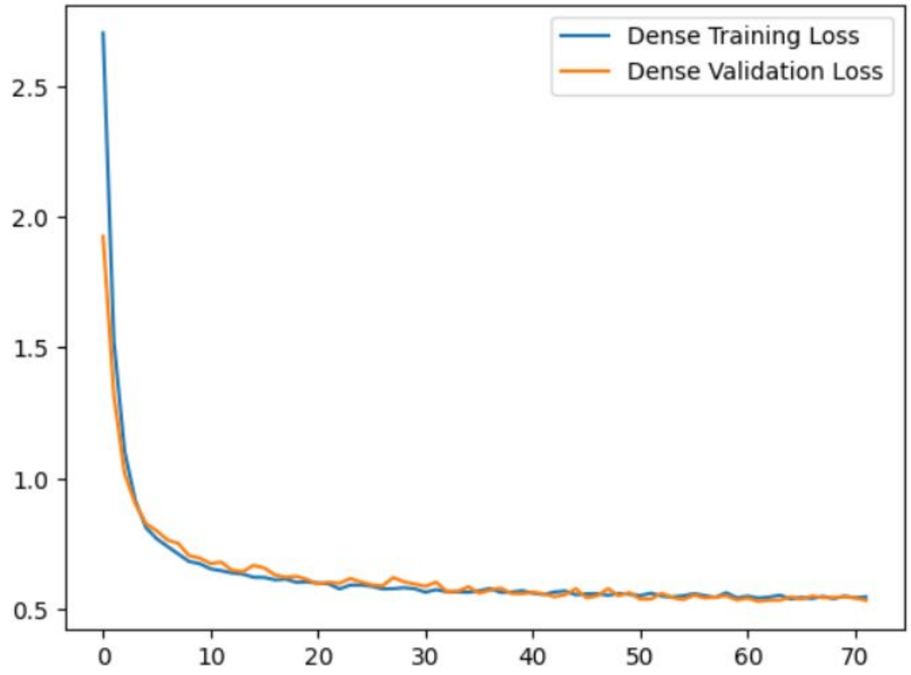
Epochs: 100

<u>Results:</u>

Mean Squared Error: 422

R^2 score for training set: 0.597

R^2 score for test set: 0.529



RNN - Long Short-Term Memory

<u>Before tuning:</u>

layers of LSTM: 2

Dropout : 0.2

Activation: relu

Optimizer: adam

Loss: mse

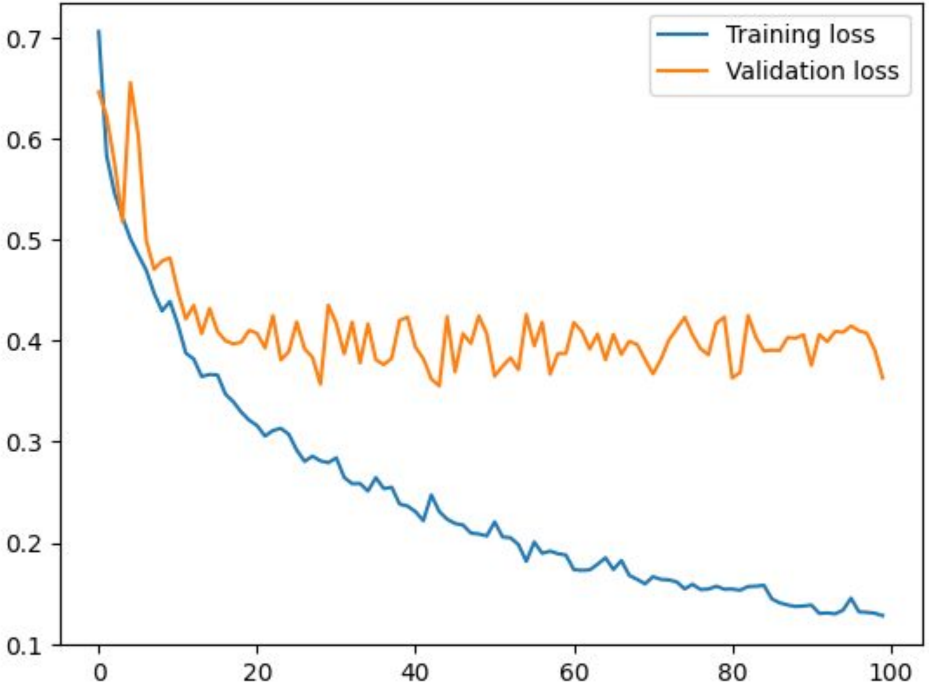
Epochs: 100

<u>Results:</u>

Mean Squared Error: 1147

R^2 score for training set: 0.881

R^2 score for test set: 0.645



RNN - Long Short-Term Memory

After tuning:

layers of LSTM: 2

Units: 512, 256

Dropout : 0.8

Activation: relu

Learning_rate: 0.0001

Loss: mse

Epochs: 100

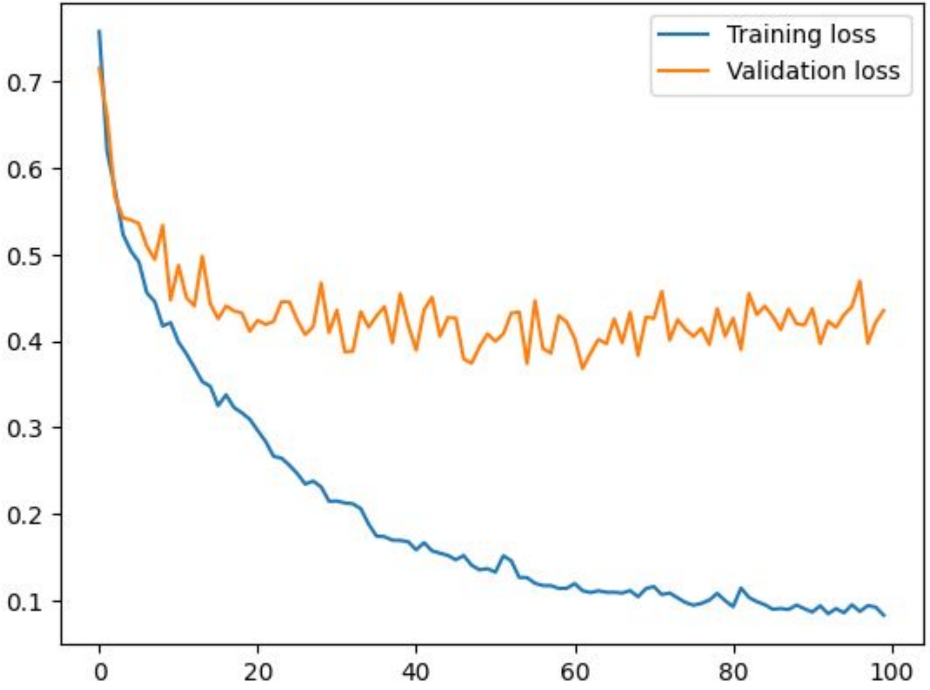
Regularization (L1): 0.0001

Results:

Mean Squared Error: 750

R^2 score for training set: 0.7571

R^2 score for test set: 0.6745



Convolutional Neural Network (1D)

<u>Before tuning:</u>

layers of Conv1D: 3

Filters: 128, 64, 32

Dropout : 0.1, 0.3, 0.0

Activation: relu

Optimizer: adam

Learning_rate: 0.0001

Loss: mse

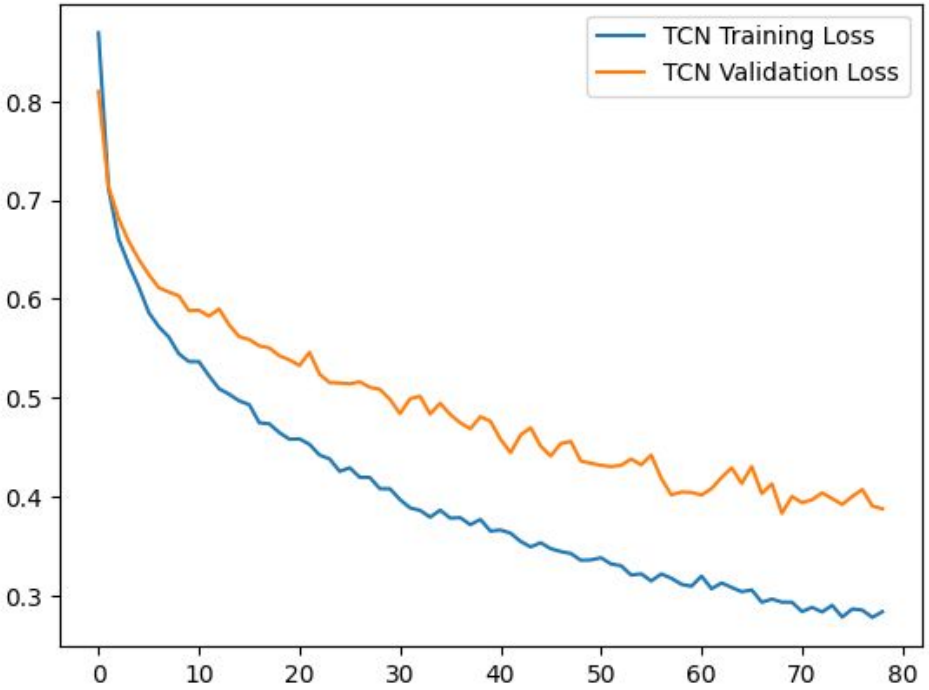
Epochs: 100

<u>Results:</u>

Mean Squared Error: 369.08

R^2 score for training set: 0.75983

R^2 score for test set: 0.64437



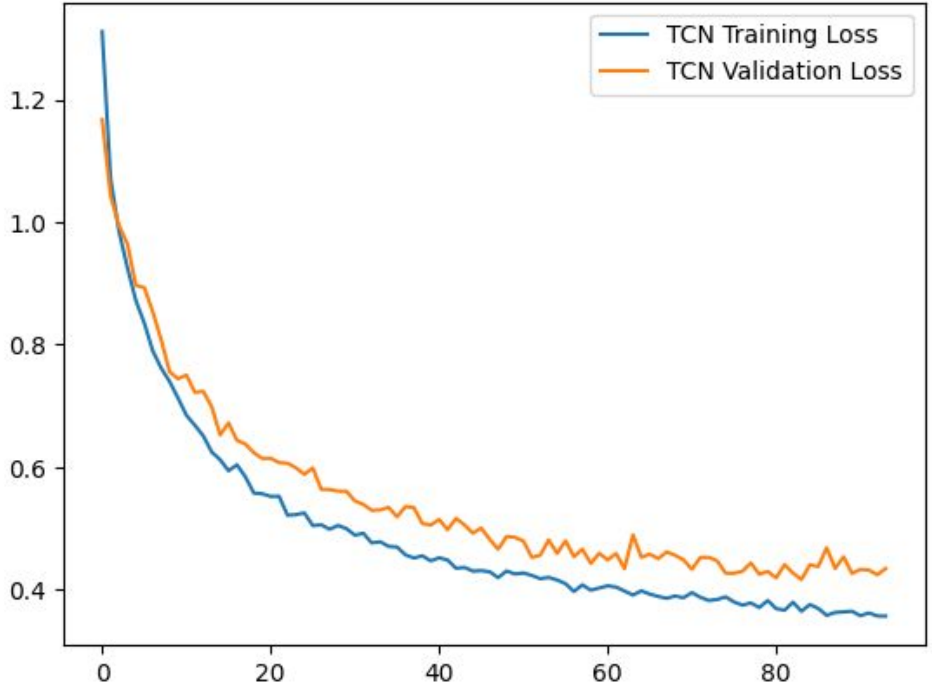
Convolutional Neural Network (1D)

<u>After tuning:</u>

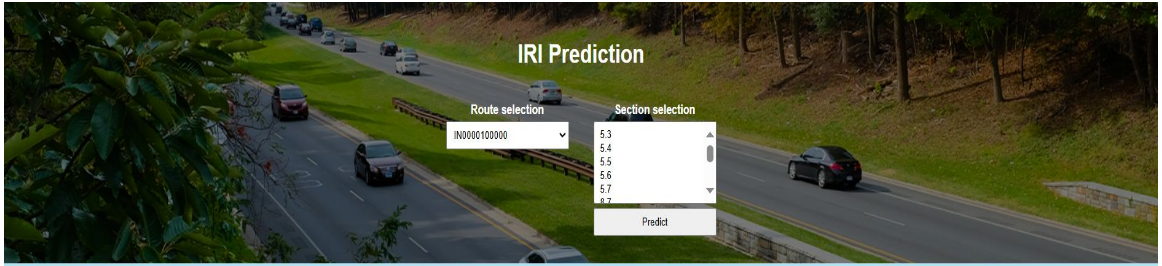
layers of Conv1D: 4
Filters: 128, 96, 96, 96
Dropout : 0.1, 0.3, 0.1, 0.1
Activation: relu
Optimizer: adam
Learning_rate: 0.00030508
Loss: mse
Epochs: 100
Regularization (L1): 0.0001

<u>Results:</u>

Mean Squared Error: 318.41
R^2 score for training set: 0.7942
R^2 score for test set: 0.70319



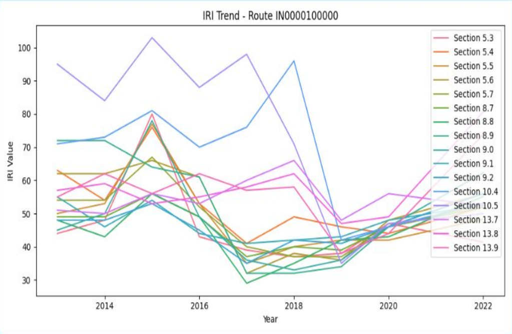
Forecasting IRI with Different Levels of Granularity



Prediction Table

ROUTE_ID	BEGIN_POIN	END_POINT	Predicted IRI
IN0000100000	5.3	5.4	41.39
IN0000100000	5.4	5.5	51.86
IN0000100000	5.5	5.6	48.32
IN0000100000	5.6	5.7	55.93
IN0000100000	5.7	5.8	52.73
IN0000100000	8.7	8.8	51.94
IN0000100000	8.8	8.9	55.15
IN0000100000	8.9	9.0	56.00
IN0000100000	9.0	9.1	62.31
IN0000100000	9.1	9.2	52.60
IN0000100000	9.2	9.3	56.00
IN0000100000	10.4	10.5	53.18
IN0000100000	10.5	10.6	49.97
IN0000100000	13.7	13.8	52.27

Graph

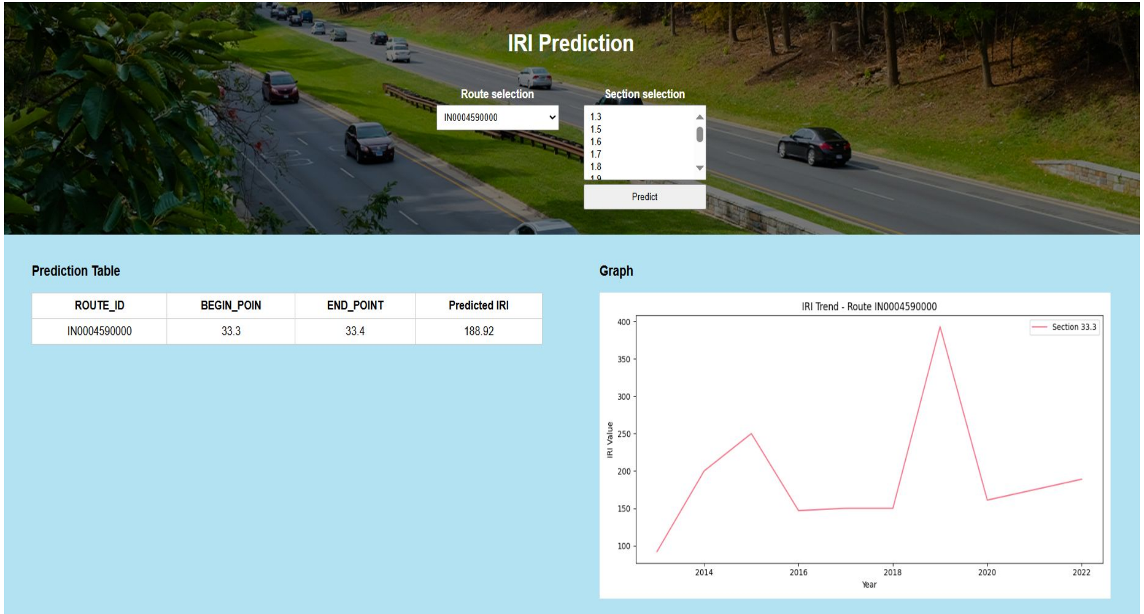


Whole route level approach.

We have create a webpage that lets user to select whole route with all its segments to get IRI predicts.

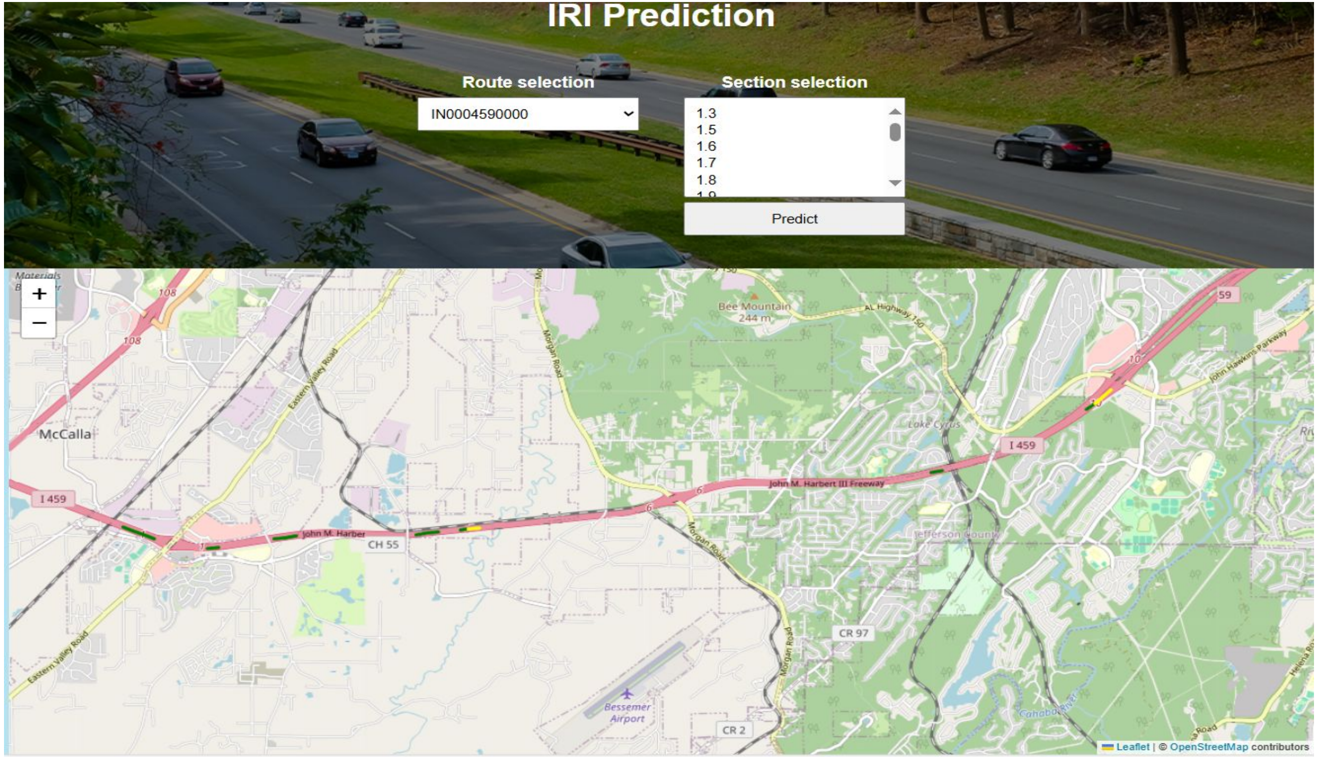
UMBC

Forecasting IRI with Different Levels of Granularity



Our interface also allows user to select specific section of a route for more granular IRI predictions

Geospatial Mapping



To identify what sections that require immediate attention. We mapped the predictions as following:

If Predicted IRI is less than 94 then the section is displayed in 'Green'.

If IRI is in between 95 to 119 the section is displayed in 'Yellow'.

If IRI is above 119 then the section is displayed as 'Red'.

Ref: [IRI Thresholds]()

Answering Research Questions

1. How does data fusion improve IRI prediction compared to single-source models?

Integrated HPMS (pavement condition) + FAF (freight flow) improved route-level accuracy.

2. What are the most influential features across models?

As we have seen in the correlation matrix we have many correlated features from both HPMS and FAF datasets such as AADT, Speed limit, tons, miles, and value..

3. How effective is the proposed model for maintenance planning?

We are going with convolutional neural networks as it produced more accuracy with less error across the dataset. Also , sequential model is helpful for predicting granular and route-level predictions.

4. Best granularity for IRI forecasting?

Using CNN predictions on both segment level and route level are optimal because of the models sequential nature.

5. Effective method for visualizing and presenting findings?

Geospatial mapping with good, moderate and need action denotations can help visualise which parts of the road requires immediate attention.

Conclusion

Our project combined

- **Descriptive analysis** (understanding current road conditions and historical patterns),
- **Diagnostic analysis** (identifying key factors like traffic and freight that drive deterioration, and recognizing why some modeling approaches failed),
- **Predictive analysis** (building a CNN model to forecast future IRI with good accuracy), and
- **Prescriptive analysis** (providing tools and visualizations to guide maintenance decisions based on those predictions).

By fusing data and applying advanced modeling, we developed a **proactive maintenance planning tool** that can help extend the **life of highways** and **optimize repair efforts**.

References

[1] (Yuanjiao Hu et al., 2022). Evaluation of pavement surface roughness performance under multi-features conditions based on optimized random forest. https://ieeexplore.ieee.org/document/9816255

[2] (Maher Mahmood et al., 2020). Multi-Types of Flexible Pavement Deterioration Prediction Models. https://ieeexplore.ieee.org/document/9122932/

[3] (Moein Latifi et al., 2021). A deep reinforcement learning model for predictive maintenance planning of road assets: Integrating LCA and LCCA. https://arxiv.org/abs/2112.12589

THANK YOU