Aalto University
MS-E2177 - Seminar on Case Studies in Operations Research

Modelling of Paved Road Deterioration

Final Report

Team Members:

Jonni Järvinen (Project manager)
Markus Junttila
Veera Wilkki
Oskar Heikkilä
Konsta Lahtinen

Returned: May 20, 2026

Contents

1 Introduction

1.1 Background

The Finnish Transport Infrastructure Agency (Väylävirasto), FTIA for short, is a Finnish government agency responsible for constructing and maintaining the public roads, railways and waterways in Finland. The road network under their management totals approximately 78 000 km, of which 53 000 km are paved roads, which carry 95% of all road traffic.

The importance of good road conditions is obvious. The general functioning of society requires huge amounts of goods to be transported and a large number of people to travel by car from point A to B safely every day. Deteriorating road conditions directly affect the safety of travel, in addition to having significant economic costs. Badly maintained roads increase the risk of accidents, force traffic to slow down, and damage cars more frequently.

The FTIA aims to mitigate these issues by regularly monitoring road conditions throughout the network, analyzing the collected data, and assigning repairs and other procedures as efficiently as possible. The goal is to make travel conditions as good as possible, for as many people as possible, while also maintaining passable travel routes everywhere.

Road conditions are currently measured with two indicators: IRI (International Roughness Index) for smoothness and URA for rut depth. Measurements are made between 1-4 years, depending on the road class and traffic volume. Other data, such as pavement type and information about past repairs, are also available. There is historical data for over 20 years.

The modelling and prediction of future road conditions is fairly difficult, and the models in use at FTIA are relatively simple and are not capable of utilizing all the historical data available. External factors that are difficult or simply cannot be measured effectively are a challenge. They include climate change, the freeze-thaw cycle every year, changes in heavy traffic capacity legislation, and regional differences in soil, weather etc. The quality of the predictions directly affects the cost efficiency of repair operations, which in turn decides the general condition and future sustainability of all roads.

The motivation for this project is the need for more accurate predictive models to model future road conditions. We aim to accomplish this by making a machine learning based model to utilize all relevant existing data.

1.2 Objectives

The main goal of our project is to make a machine learning based model to predict the behaviour of URA values for all roads in the dataset as accurately as possible. The potential secondary goal would then be to assess the most damaged continuous sections for repair.

The dataset of 20+ years from FTIA has a variety of different measurement categories. The main variables to consider are previous measurements of IRI and URA, volume of regular

and heavy traffic, pavement structure and maintenance history. However, the amount of data in each category varies greatly between different road sections, and we know some data to be inaccurate or missing completely. This has to be taken into account when building the model.

Because we are making a new model, rather than improving upon an existing one, there is a need to try and compare between different approaches. Possible candidates would be tree-based ML models, like Random Forest, Gradient Boosting or XGBoost, or time-series models.

Finally, we have been asked to analyze, whether climate change, regional variation, or the 2013 76-ton reform for vehicle combinations have any observable effects in road deterioration.

We seek to deliver a comprehensive and accurate model to help FTIA make informed and efficient pavement management decisions in the future.

2 Literature Review

2.1 Traditional approaches

Traditional road deterioration modelling techniques can be divided into deterministic and probabilistic models. Deterministic models, often divided into mechanistic, empirical and mechanistic-empirical models, aim to predict road condition values directly, using mathematical functions [1]. Typical examples of deterministic models include linear and non-linear regression. In contrast, probabilistic models estimate future road condition or life expectancy through probabilities. The most widely used probabilistic model in road deterioration modelling is the Markov chain, but other methods, such as Bayesian analysis, are also used [1].

While traditional deterministic and probabilistic models may yield accurate results in certain contexts, they have considerable limitations. Many deterministic models tend to oversimplify or ignore complex relationships between variables in order to remain interpretable, at the cost of predictive accuracy [1]. Similarly, traditional probabilistic models make assumptions about underlying probability distributions, which may not accurately represent real-world conditions [1]. In recent years, machine learning (ML) methods have emerged as a promising alternative, as they are capable of capturing complex relationships without relying on predetermined functions or probability distributions [1].

2.2 Machine learning methods

Supervised machine learning methods have been shown to outperform traditional approaches for predicting both IRI and rutting values [2]. These methods include artificial neural networks (ANNs), random forests and gradient boosting, among others. ML approaches for pavement

performance prediction can broadly be divided into static and forecasting models. In static models, ML methods are used to estimate current values of IRI or rut depth using other road condition variables. In forecasting models, the aim is to predict future values of IRI or rut depth using historical data. Reviewing both approaches provides valuable insights into which ML models best capture the complex relationships that characterise pavement performance data.

ANNs are one of the most widely applied ML methods for pavement performance prediction in literature [2]. For example, Choi & Do [3] use a recurrent neural network (RNN) algorithm to predict pavement conditions of Korean road segments, using data from the previous 10 years. The pavement condition is represented by variables such as cracking, rut depth and IRI. Their RNN algorithm predicts pavement deterioration for the different road segments with a high coefficient of determination (R^2) of 0.71-0.87.

Abdualaziz Ali et al. [4] compare a multiple linear regression (MLR) model to an ANN model for predicting IRI based on other pavement distress variables, such as pavement age, rutting and cracking. They use data from two climate regions in the United States and Canada: wet freeze and wet no freeze. Particularly the wet freeze regions may be comparable to certain climate regions in Finland. Their results show that while both models predict IRI quite accurately, ANN is the higher performing model with a remarkably high R^2 of 99.1% for wet freeze regions, and R^2 of 97.5% for wet no freeze regions.

Random forests have also been proven to be a high performing method for road deterioration modelling [5] [6] [7]. Marcelino et al. [5] predict IRI for 5- and 10-year horizons using a random forest algorithm, with time series data from the United States and Canada (LTPP database). More precisely, the dataset they use contains previous IRI values, along with structural, climatic, and traffic variables. Both models perform very well: the 5-year prediction model's R^2 value is 0.98, and the 10-year prediction model's R^2 is 0.93. The models also show strong generalisation capability across different road types.

Naseri et al. [6] use random forest as one of five ML algorithms to test a novel feature-selection method for IRI prediction. The four other tested models are: support vector machine, multi-layer perceptron (a type of ANN), decision tree regression, and multiple linear regression. The highest accuracy for IRI prediction is obtained using the random forest algorithm, with a testing R^2 value of 0.945 on average across the different feature selection methods.

In a master's thesis, Hasan [7] predicts pavement rutting using four different ML methods: decision tree, gradient boosting, ANN, and random forest. Most notably, the dataset used in the thesis is the same as the one we use in this project. The results show that random forest outperforms all other tested models with rut depth RMSE of 1.76 mm, rut depth MAE of 1.17 mm, and an R^2 score of 0.81.

While ANNs and random forest algorithms have yielded accurate predictions for road deterioration modelling, gradient boosting has shown competitive or superior performance to both methods in numerous recent studies [8] [9]. Adnan & Erfani [8] predict IRI over 2- and 3-year horizons under both maintenance and no-maintenance scenarios. They compare ANNs and random forest to two gradient boosting algorithms: XGBoost and CatBoost. The highest performing model overall is CatBoost with R^2 values ranging between 0.790 and 0.918 depending on the time horizon, the maintenance scenario, and the parameter scaling method.

Liu et al. [9] use Bayesian-optimized Natural Gradient Boosting (BO-NGBoost), a probabilistic ML method, to predict IRI in cold climates using data from the United States and Canada (LTPP database). The algorithm is compared to ANNs, random forest, and XGBoost and is proven to outperform all of them with an R^2 value of 0.897. Given its probabilistic nature, BO-NGBoost is able to quantify uncertainty directly, unlike the deterministic ML models it is compared to.

Chen et al. [10] compare random forest and XGBoost for predicting pavement rutting using data from RIOHTrack, a full-scale track test pavement in China. They also examine the effects of using the snake optimization algorithm for hyperparameter tuning. Their results show that XGBoost is the superior model with and without optimized hyperparameter tuning, achieving remarkably high testing R^2 values of 0.951 and 0.995, respectively.

Liu et al. [11] use XGBoost to predict pavement rutting and cracking using traffic, material and climate variables, among others. Their study shows impressive performance for rutting prediction, achieving an R^2 value of 0.91. XGBoost is also shown to outperform support vector machines, decision tree, random forest, and K-Nearest Neighbor methods.

Based on the studies reviewed above, gradient boosting methods have been shown to be particularly high-performing for both IRI and rutting prediction. Random forest is also competitive, particularly when using the same dataset as in this project [7]. These results motivate the use of these two modelling methods in particular.

2.3 Key predictors

Many recent pavement performance modelling studies, such as [8], [9] and [11], perform SHAP analysis to extract feature importance information for the models they use. SHAP (SHapley Additive exPlanations) is a method first introduced by Lundberg & Lee [12] that is used to improve interpretability of complex machine learning and AI models. In the present context, it helps identify parameters that affect road deterioration the most, which in turn informs road management decisions.

Applying SHAP, both Adnan & Erfani [8] and Liu et al. [9] find that the previous IRI value is the dominant factor in predicting future IRI values. Liu et al. [11] finds that parameters related to load and environmental conditions, such as temperature and truck volume, are the

most significant factors for rutting prediction. All of the aforementioned studies use gradient boosting as their prediction method.

Other studies relying on built-in feature importance metrics or sensitivity analysis largely confirm these findings. Marcelino et al. [5] and Naseri et al. [6] both find that previous IRI values are the key predictors for future values. Naseri et al. [6] also find pavement age to be significant, which aligns with a sensitivity analysis performed by Abdualaziz Ali et al. [4].

Chen et al. [10] find that the number of shaft loads is the most important feature for rutting prediction when using random forest for modelling, which confirms the conclusion made by Liu et al. [11] on the importance of traffic load variables. Hasan [7], who employs the same road dataset used in this project, finds that pavement age is the most important factor for rutting prediction, while traffic volume indicators also play a significant role.

3 Data

3.1 Data sources and structure

The dataset is based on historical road condition and maintenance data provided by the Finnish Transport Infrastructure Agency. The data spans 26 years of observations and the primary raw dataset contains measurements, and maintenance operations, covering over one million segments and several million observations in total.

Originally, the data was structured in a wide format, where each row corresponded to a single road segment and historical measurements were stored in index columns. These included pavement condition indicators such as the International Roughness Index (IRI) and rut depth, along with time stamps and maintenance-related variables. This format is efficient for storing the data. However, it does not clearly represent how observations evolve over time and is therefore not directly suitable for training machine learning models.

3.2 Data cleaning and filtering

Several preprocessing steps were applied to ensure data quality and consistency. Observations were filtered to include only valid measurements within realistic bounds, and only data from recent decades were retained. Rows without a valid future target were excluded from the supervised dataset.

To further assess data quality and justify the applied filtering criteria, the distributions of IRI and URA values were examined, as shown in Figure 1. Both variables exhibit right-skewed distributions with a small number of extreme values forming long tails. To improve model stability and reduce the impact of outliers, upper bounds were applied to both variables. For IRI, values above 10 were removed, while for URA, values above 40 were excluded from the dataset

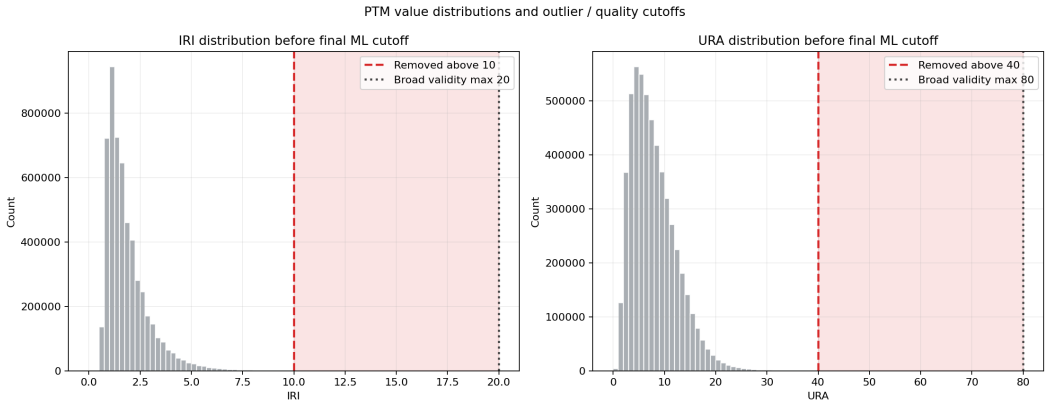

Figure 1: Distributions of IRI and URA values before filtering. The dashed lines indicate the thresholds used to remove extreme values for machine learning.

Figure 1 illustrates that the majority of observations lie well within the selected thresholds, while extreme values occur infrequently but could disproportionately affect model training. Removing these outliers helps ensure that the model focuses on realistic road condition behavior and improves the robustness of the predictions.

3.3 Event-based data transformation

To enable training machine learning models, the raw data was transformed into a different structural format. The first major transformation step converted the wide-format into an event-based representation. In this representation, each observation corresponds to a single event occurring at specific time for given segment. If both maintenance and measurement occured on the same day, maintenance was assumed to happen first, so that the measured condition reflects the effect of the intervention.

During this transformation, several other features were constructed, including differences between consecutive measurements, lagged condition values, and counts of maintenance actions between measurement events. Additionally, pavement lifecycles were interpreted based on changes in condition variables. Significant improvements in rut depth were interpreted as major resets, which mark the beginning of a new lifecycle, while smaller interventions were accumulated as part of the lifecycle history.

This event-based representation enabled the dataset to capture the evolution of road condition and the impact of maintenance over time.

Figure 2 shows an example of URA evolution within a single pavement lifecycle. The shaded region indicates the selected lifecycle, bounded by inferred lifecycle start and end points. Blue points represent condition measurements within the lifecycle, while gray points show measurements outside the selected lifecycle. The figure illustrates how pavement condition evolves over time and how lifecycle boundaries are defined based on changes in condition.

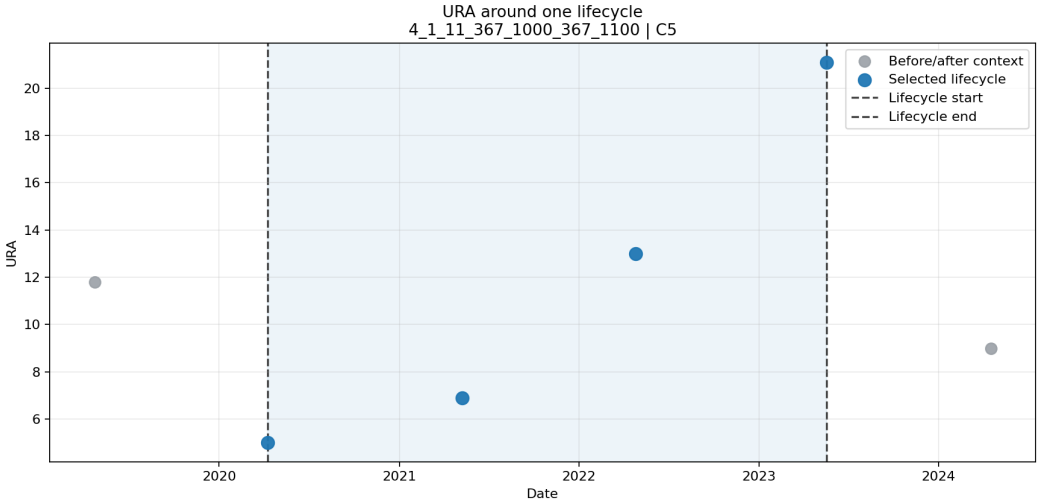

Figure 2: An example of URA evolution within a single pavement lifecycle.

3.4 Modeling dataset construction

From the event-based dataset, a supervised modeling dataset was constructed. In this dataset, the unit of observation is a single pavement condition measurement. Each row was enriched with historical and contextual features derived from the event history.

As a result, each observation represents the state of a road segment at a given time, while the target corresponds to its condition at a later time. The task can therefore be interpreted as a forecasting problem with variable time horizons, ranging from approximately one to four years.

Only measurement events were included in the final supervised dataset, while maintenance events were used for feature construction. In addition, observations without a valid future target were removed, ensuring that each row has a well-defined prediction objective.

3.5 Feature sets

Many predictor variables were used in the modeling process. They represent multiple aspects of road condition dynamics. Table 1 contains the features used for modeling, along with their meanings.

Table 1: Feature descriptions.

Feature	Meaning	Usual value range
URA	Current rutting	$[0, 40]$
IRI	Current roughness	$[0, 5]$
target_horizon_years	Years from current PTM[1] to prediction date	$\{0, 1, ..., 7\}$
KVL	Traffic volume in both directions	$[0, 60\,000]$
KVL_raskas	Heavy traffic volume	$[0, 3\,000]$
KVL_kaista	Traffic volume in the specific direction	$[0, 30\,000]$
Nopeus	Speed limit	$\{10, 20, ..., 80, 100, 120\}$
Toim_lk	Functional road class	$\{1, 2, 3\}$
prev_URA, prev_IRI	Previous PTM[1] condition value	$[0, 40]$, $[0, 1]$
Delta_t_years	Years since previous PTM[1]	$\{0, 1, ..., 7\}$
observed_lifecycle_age_years	Years since the start of the inferred lifecycle	$\{0, 1, ..., 10\}$
Minor_TP_Count	Cumulative minor-treatment count in inferred lifecycle	$\{0, 1, ..., 10\}$
tp_count_interval	TP[2] count since previous PTM[1]	$\{0, 1, 2\}$
surface_type_current	Latest known surface category on the segment	{None, AB, SMA, PAB-V, Sora}
material_type_current	Latest known material/work category on the segment	{None, LTA, REM, MP, MPKJ}
years_since_material_update	Years since the TP[2] event providing current material/surface info	{0,1,...,20}

[1] PTM = pavement condition measurement (päällysteiden palvelutasomittaus)

[2] TP = maintenance (toimenpide)

The most important inputs were the current condition variables, including IRI and URA, which provided a baseline for the prediction. Seasonality aspects were captured through different lagged variables, such as previous condition values (prev_IRI and prev_URA) and elapsed time since previous measurement (Delta_t_years).

Traffic-related variables were used to represent the road usage intensity. These included for example traffic volume (KVL and KVL_kaista) and heavy traffic share (KVL_raskas). Maintenance information was incorporated through both interval-level indicators and lifecycle-level summaries describing the history of interventions.

Road's material state was also derived from maintenance events. These variables describe the current surface type, material type, and the time since the most recent material update.

Additional contextual information was provided by static road characteristics, such as speed limits (`Nopeus`) and road classification (`Toim_lk`), as well as lifecycle-related variables that describe the state of the road within its current maintenance cycle.

Figure 3 shows the five different feature sets that were tested: *current*, *current_static*, *current_static_lag*, *current_static_lag_lifecycle*, and *current_static_lag_lifecycle_material*. The *current* set contains three features, containing the bare minimum information for prediction, i.e., current road condition values along with the desired prediction horizon. The second feature set adds static road properties, such as traffic volume and speed limit. The *current_static_lag* set adds temporal lag features, and the *current_static_lag_lifecycle* set also includes lifecycle context. Finally, the largest set with 17 features in total, contains all of the previous information along with material and surface history.

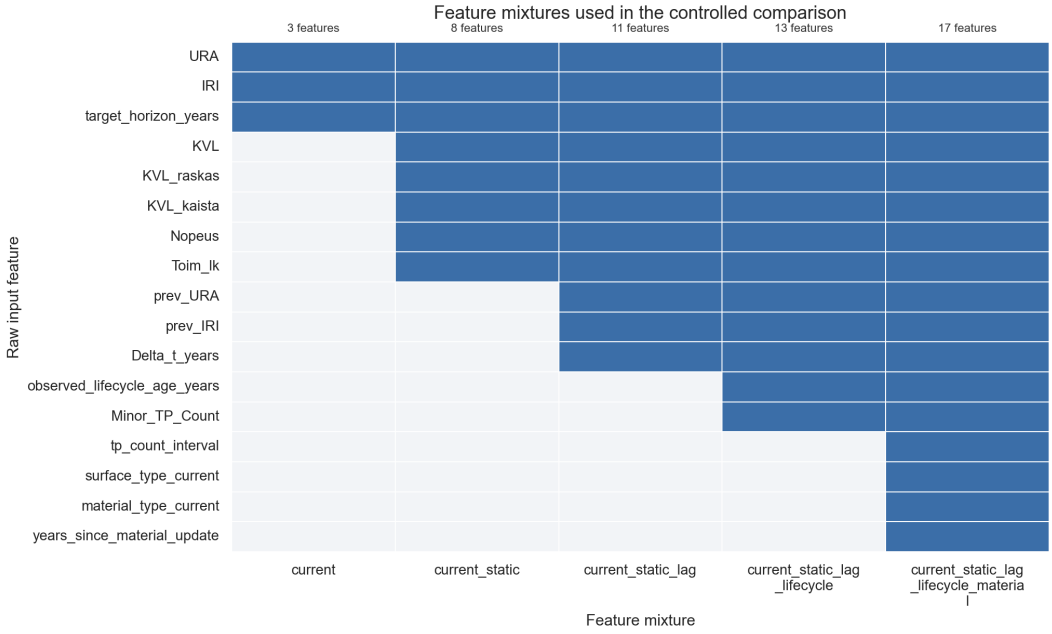

Figure 3: Feature sets used for model comparison.

4 Methods

Five different modeling methods were tested: linear regression, ridge regression, random forest, gradient boosting, and histogram gradient boosting. Linear regression and ridge regression are traditional statistical models, that were tested to establish whether the added complexity of machine learning models is justified. Random forest and gradient boosting methods were chosen based on the literature review, which showed these machine learning methods to be particularly high-performing in road deterioration modeling. All aforementioned models were compared to a persistence baseline model, i.e., a model that assumes road condition values to remain unchanged.

4.1 Linear regression

One-dimensional linear regression attempts to fit a straight line to the training data to best represent the target variable often using the least-squares method, i.e., trying to minimize the sum of squared distances between each data point and the output of the model. With multiple predictor variables the target variable is predicted by fitting a hyperplane to the training data. [13]

Figure 4 shows an example of a linear regression model applied to the television advertising budget (TV) to predict the sales of a given product (Sales) [13]. The red dots represent the data points and the blue line represents the fitted regression line.

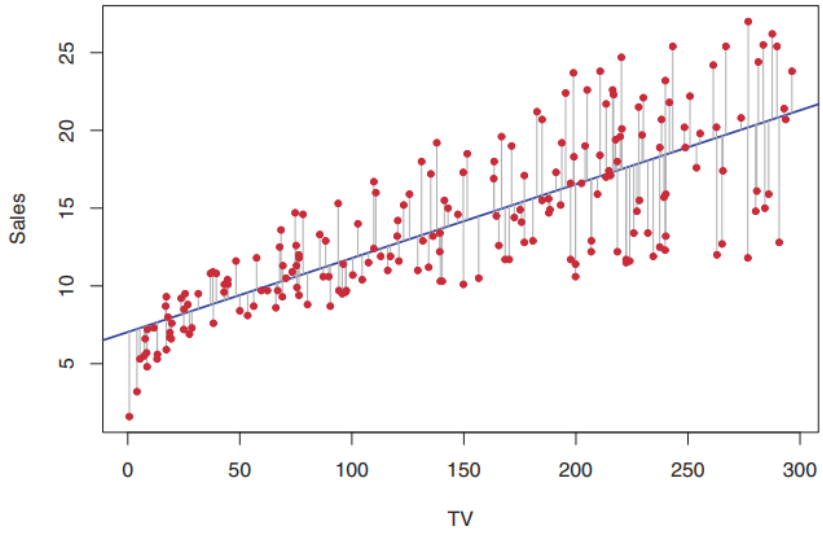

Figure 4: Illustration of linear regression. Source: [13].

In order to avoid overfitting to the training data, regularization is often applied by adding a penalty term to the loss function being minimized. Ridge regression is a regularized version of linear regression that uses L2 regularization, which penalizes large coefficient values by adding the sum of squared coefficients to the loss function. [13]

4.2 Base learner models

4.2.1 Decision trees

Methods such as Random Forest rely on creating decision trees by partitioning the feature space into regions. An example of a partition from our data could be $P_1 = \{X \mid \text{Speed} >= 80\}$ and $P_2 = \{X \mid \text{Speed} < 80\}$. In this partition, the first region includes all road segments with a speed limit greater than or equal to 80 km/h and the second includes those with a speed limit strictly less than 80 km/h. These regions are then further partitioned with different features until the method reaches a sufficiently large tree whose nodes are the partitioned regions. [13]

4.2.2 Bagging

The decision trees formed in the previous subsection suffer from high variance, i.e., splitting the feature space into two subsets will most likely yield significantly different decision trees. Because of this, *bootstrap aggregation* or *bagging* is used to normalize the trees. [13]

Bagging is an ensemble machine learning method meaning that it utilizes multiple simple models to boost performance. It relies on the fact that, given independent identically distributed observations $Z_1, ..., Z_n$ each with variance σ^2, the variance of the mean $\overline{Z}$ is given by σ^2/n, i.e., the variance of the mean tends to decrease as the number of observations grows. This means that the variance of the prediction model can be decreased by taking the average of multiple decision trees built from different subsets. Taking multiple different samples of a given data set is called *bootstrapping*. [13]

4.3 Random Forest

The *Random Forest* method resembles bagging but for each decision tree only a small, random sample of m predictors is used, typically such that $m \approx \sqrt{p}$, that is, each decision tree considers only m predictors such that m is approximately equal to the square root of the number of total predictors. [13]

This method ensures that, even in the case of one predictor dominating the data set, the bagged trees will be significantly less correlated with each other since most do not consider this strong predictor. With the random forest method, one obtains a bagged tree which will substantially reduce the variance compared to a single tree. [13]

4.4 Gradient Boosting

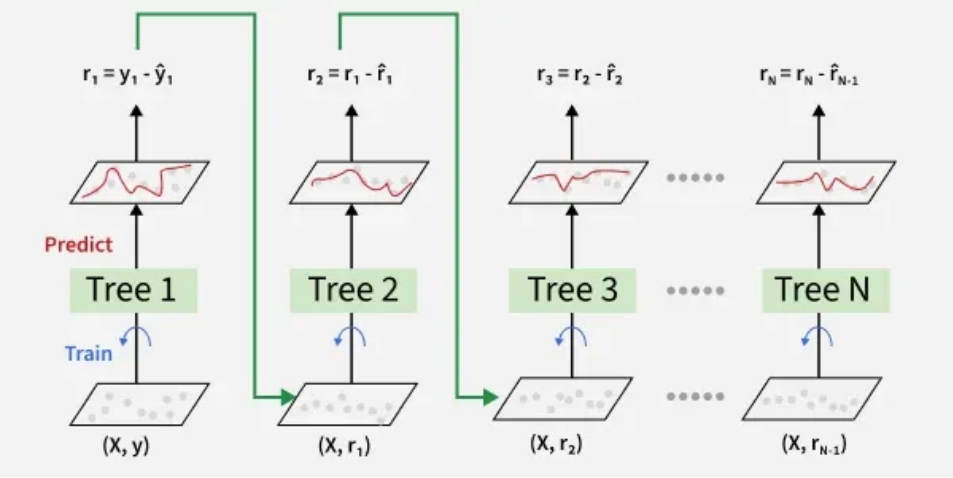

Figure 5: Gradient Boosting visualized. Source: [14]

Gradient Boosting is illustrated in Figure 5. It starts with a small initial decision tree with random predictions, called a weak learner. For this tree, the method calculates its error by subtracting the predictions from the data to evaluate its performance with respect to the real data. [15]

After this, the model trains a new weak learner to predict the errors given by the previous weak learner and tests it. If this model performs better than a random guess, it will be multiplied by a learning rate and included in the ensemble. The learning rate is between 0 and 1 and depends on the performance of the weak learner. This step is repeated and the performance of the overall ensemble is monitored on the validation set until the results are satisfactory. [15]

4.5 Model selection

From the literature review, Random Forest and Gradient Boosting were selected as candidate methods based on the historical performance in similar problems. Histogram Gradient Boosting was also added to the group, as a more sophisticated version of Gradient Boosting, suitable for large datasets. After discussing the project, its goals and intermediary model results with the FTIA, the objective to predict IRI and URA values was changed to prediciting only URA values. IRI was deemed too difficult to model, while having a comparatively small impact on the decision making framework at FTIA that is in the scope of this project.

In model comparison, Random Forest, Gradient Boost and Histogram Gradient Boost models were compared to select the best performing one. Ridge Regression, Linear Regression and Persistence models were also added to the comparison as simple baseline models to give a point of comparison to the main models.

Different feature groups (shown in Figure 3) were used for training the models in order to find the best performing feature set for each model. All of the models performed best with the split including all of the features, shown in Figure 6. This indicates that every feature is significant for the prediction of URA and IRI values and also makes the models very easily comparable.

In model training, the dataset was split with 70%, 15% and 15% into training, validation and testing sets, respectively. Because the dataset was constructed with each datapoint representing a single event, the split had to take into account that all of the datapoints associated with a single road segment are contained in a single split. This ensures that any information considering a single road segment used in training does not contaminate the testing or validation sets, which could lead to the model being overly optimistic. The difference in ELY regions was also taken into account, making sure that the split of 70/15/15 also applies considering each individual ELY region. Different regions have different dominating road conditions, and ensuring equal partitioning for all of the regions prevents the skewed distribution of any single region from dominating any of the splits.

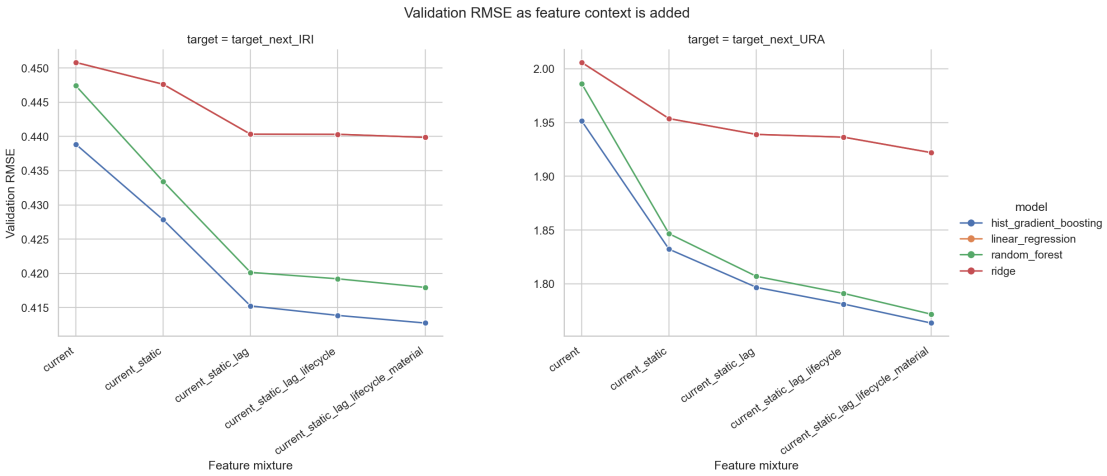

Figure 6: Validation set URA RMSE values for different model candidates and feature sets.

The models were tested with the following metrics on the test set:
- Mean absolute error (MAE): Mean of the errors i.e. differences between predicted values and target values.
- Root mean squared error (RMSE): Square root of the mean of the squared errors.
- Coefficient of determination (R^2): Measures the proportion of variation in the data which is explained by the model.

As can be seen from Table 2, Histogram Gradient Boosting is clearly the best model, with RMSE of 1.7652 and R^2 of 0.8408, showing significant prediction accuracy over the baseline Persistence model with RMSE of 2.6460 and R^2 of 0.6422. Random Forest is a close second, with 1.7742 RMSE and 0.8391 R^2. Normal Gradient Boosting had noticeably worse results, with RMSE of 1.8412 and R^2 of 0.8268. Interestingly, even the most basic models of Linear and Ridge Regression improve significantly upon the Persistence baseline, coming somewhat near the best models, both having RMSE of 1.9235 and R^2 of 0.8109. This indicates that there is a very strong linear relationship between the explanatory and target variables.

Based on these results, HGB was selected as the best architecture to build the final model. For the implementation of the final model, further improvements were made to the base model. It was noticed that changing the target variable from the next URA value to the change between the current and the next URA values resulted in a slight improvement in performance. From Table 3 we can see that this adjustment reduced RMSE from 1.7652 to 1.7542 and increased R^2 from 0.8408 to 0.8427. Lastly, the HGB delta model went through hyperparameter fine tuning, resulting in the final version of the prediction model.

Table 2: Performance comparison of different models. Values are reported for the validation set.

Model	**Target Type**	**MAE**	**RMSE**	R^2
HGB	direct	1.1506	1.7652	0.8408
Random forest	direct	1.1411	1.7742	0.8391
Gradient boosting	direct	1.2093	1.8412	0.8268
Linear regression	direct	1.2864	1.9235	0.8109
Ridge	direct	1.2864	1.9235	0.8109
Persistence	direct	1.8945	2.6460	0.6422

Table 3: HGB target variable comparison

Target Type	**Test MAE**	**Test RMSE**	**Test R^2**
HGB direct	1.1506	1.7652	0.8408
HGB delta converted to actual	1.1446	1.7542	0.8427

5 Results

The results of the project are reported for a histogram gradient boosting model, that was obtained through the model selection and hyperparameter tuning process. The model was trained with all available candidate features. The final overall results that were obtained on the held out test set are shown in Table 4. The model achieves rut depth RMSE of 1.64 mm, MAE of 1.06 mm, and R^2 value of 0.862.

Table 4: Final test set results for the selected histogram gradient boosting model.

Metric	Test set result
RMSE	1.64 mm
MAE	1.06 mm
R^2	0.862

The obtained test-set performance is competitive with previous machine-learning studies on rutting prediction. The most directly comparable benchmark is [7], that studied pavement rutting prediction using Finnish pavement data and reported a best model performance of $R^2 = 0.81$, MAE of 1.17, and MSE of 3.09. In comparison, the selected histogram gradient boosting model in this project achieved better MAE and R^2 values, suggesting that the event-based data transformation, feature construction, and model selection process improved predictive accuracy.

Compared with studies based on other datasets, the result is also within the range reported in recent rutting-prediction literature. Liu et al. [11] reported an R^2 value of 0.90 for rutting prediction using XGBoost on LTPP data. Alnaqbi et al. [16] reported a best selected-feature rut-depth model with RMSE of 1.9371, MAE of 1.3308, and R^2 value of 0.71 using the LTPP database. Higher values have been reported in controlled full-scale test-track settings: for example, Cheng et al. [17] reported test-set R^2 values of 0.9767 for random forest, 0.9835 for gradient boosting decision trees, and 0.9740 for Extra Trees using RIOHTrack data. These values should not be interpreted as directly comparable to the present national-network forecasting problem, since controlled track datasets have more consistent loading, materials, and measurement conditions than heterogeneous road-network data.

Figure 7 shows rut depth RMSE values against time horizons of 1 year, 2 years, 3 years and 4 years. It can be seen in the figure, that the model achieves significantly better results when making predictions for shorter time horizons. This result is not surprising, as the variance of the rut depth change on shorter horizons is likely smaller than on longer horizons. The used dataset also contains much more data for shorter horizons and the short horizon data is likely of better quality, which makes shorter horizons easier for the model to predict.

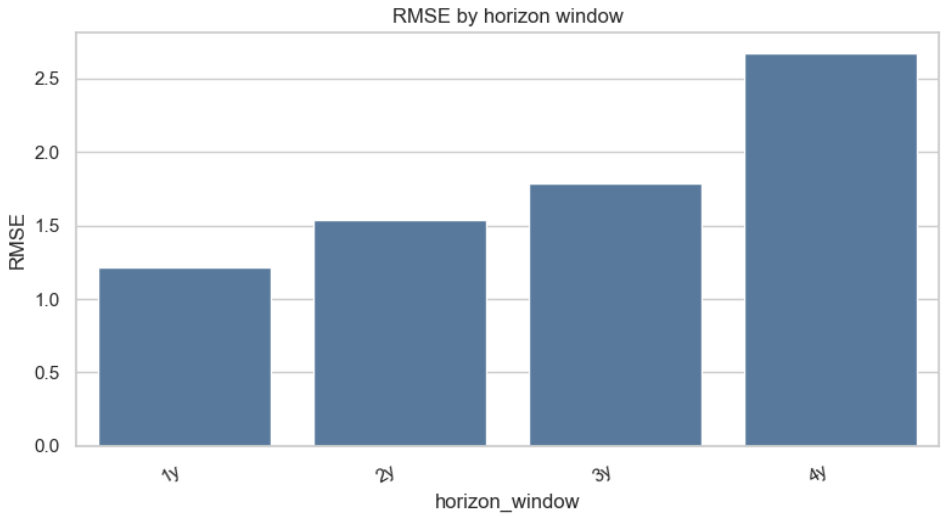

Figure 7: Rut depth RMSE by horizon window. The horizon windows are not exact, but are determined in the following way: 1-year: 274 to 457 days, 2-year: 639 to 822 days, 3-year: 1004 to 1187 days, 4-year: 1370 to 1553 days.

Figure 8 shows the rut depth RMSE that the model achieves for test data coming from different ELY centres of Finland. It can be seen that the model performs clearly better on data that originates from southern Finland, and worse on data that originates from northern Finland. The performance distribution is expected to look like this, as the road network is probably better looked after and maintained in more densely populated areas, such as southern Finland.

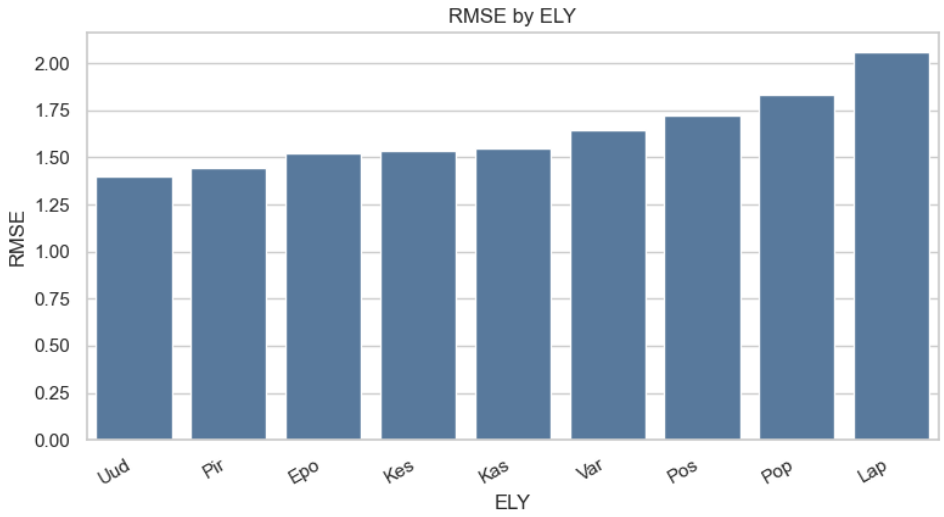

Figure 8: Rut depth RMSE by ELY area. From left to right, the abbreviations correspond to Uusimaa, Pirkanmaa, Etelä-Pohjanmaa, Keski-Suomi, Kaakkois-Suomi, Varsinais-Suomi, Pohjois-Savo, Pohjois-Pohjanmaa, and Lappi.

5.1 Forecasted poor-condition road length

As an applied use case for the final model, we estimated how much road length is expected to newly become poor-condition road during the next four years. In the Finnish rut-depth classification, condition classes are denoted by KL values [18], and KL2 is the first class considered poor condition in this analysis. For this analysis, the latest available measurement was selected for each 100 m road segment. Segments with `KVL` below 350 were excluded, since they are outside the rut-depth target scope. For the remaining segments, the KL2 lower boundary was assigned based on the official rut-depth threshold table using traffic volume and speed limit. The final URA model was then used to predict rut depth after 1, 2, 3 and 4 years.

The analysis covered 682 729 road segments, corresponding to 68 273 km of road length. At the latest measurement, 3 055 km were already at KL2 level or worse. Among the segments that were still below the KL2 boundary, the model predicted that 8 247 km would newly cross the boundary within four years. The yearly first-crossing estimates are shown in Table 5. Each segment is counted only in the first year where its predicted rut depth reaches the KL2 boundary.

Table 5: Forecasted first crossings of the KL2 rut-depth boundary.

Year	New 100 m segments	New road length (km)	Cumulative road length (km)
1	25 661	2 566	2 566
2	33 807	3 381	5 947
3	14 181	1 418	7 365
4	8 821	882	8 247

The result suggests that a substantial share of currently acceptable road length is close enough to the KL2 boundary to become poor within a short forecast horizon. The estimate should be interpreted as a static deterioration forecast: it does not account for future maintenance, resurfacing or reconstruction actions that would prevent some segments from crossing the boundary in practice.

5.2 SHAP analysis

SHAP (SHapley Additive exPlanations) is a method that is used to quantify how much each input feature affects the model's output for a single observation [12]. The method is based on cooperative game theory: each feature is considered as a "player" in a "game" who works together with other "players" to obtain a "payout", i.e., predict a certain value [19]. The SHAP value of a single feature is defined as its average marginal contribution to the prediction across all possible feature subsets, i.e., "player coalitions" [19].

Feature importance can be examined directly using the mean absolute SHAP value, where higher values indicate a larger average impact on model predictions. We calculated SHAP values for random subsets of 2000, 5000, and 10 000 observations, and noticed that the feature importance rankings remained stable across all sizes. Figure 9 shows the mean absolute SHAP values of each feature for the sample of 10 000 observations.

Figure 9 shows that `target_horizon_years`, which represents how far in the future we are predicting rutting values for, is by far the most influential feature with a mean absolute SHAP value of over 0.50. This is not unexpected, given that an increase in prediction horizon naturally makes rutting predictions larger.

The second most influential feature is `KVL_kaista`, which represents lane traffic volume. This result makes sense intuitively: high traffic volume accelerates rutting. It is also consistent with results presented in Section 2.3, where traffic load variables were found to be key predictors of rutting.

Based on Figure 9, the previous rutting value, i.e., `prev_URA`, is the third most influential feature for predicting future rutting. However, its effect is surprisingly small compared to the previous two, with a mean absolute SHAP value of ≈ 0.33. After `prev_URA`, feature

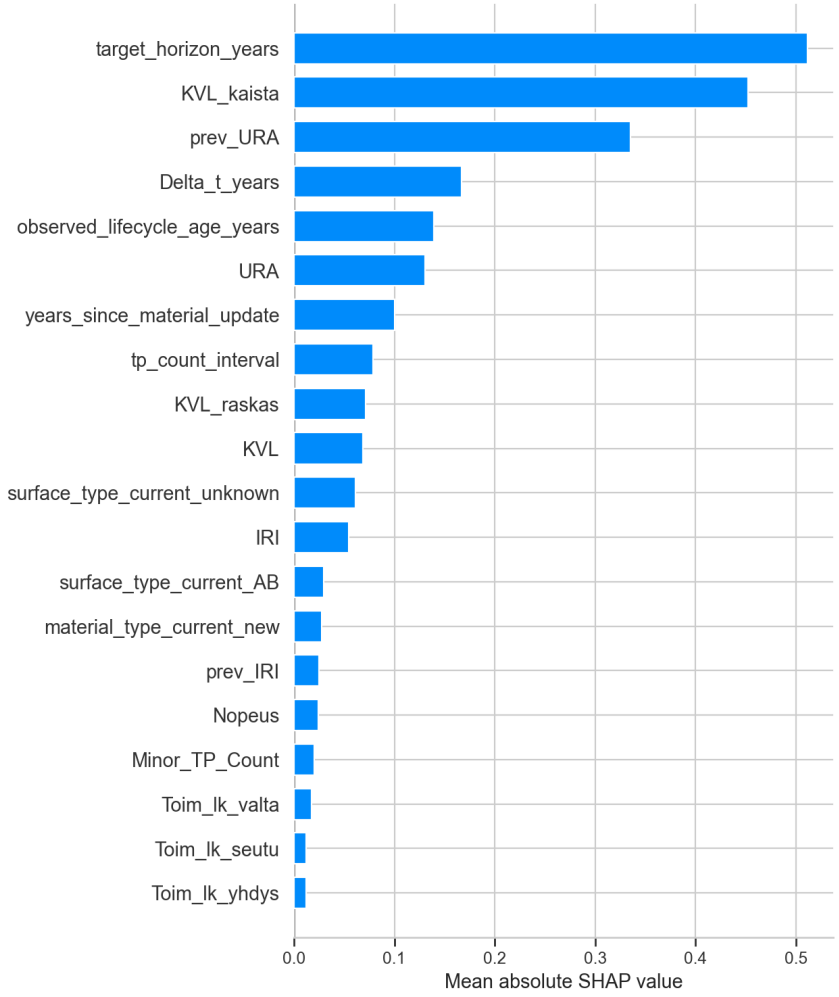

Figure 9: Mean absolute SHAP values of all features for a random sample of 10 000 observations.

importance becomes considerably smaller: mean absolute SHAP values are all smaller than 0.2.

6 Discussion

The results of this project indicate that machine learning methods are suitable for modelling pavement rutting behaviour with relatively high accuracy using historical road condition and maintenance data. In particular, the selected tree-based models were able to capture complex relationships between traffic, maintenance history, and pavement condition variables. These results were consistent with the findings presented in the literature review.

One of the main strengths of the project was the event-based transformation pipeline developed for the raw FTIA data. The original dataset structure was not directly suitable for

machine learning applications, and therefore preprocessing and feature engineering were required before meaningful modelling could be performed.

However, the project had also several limitations. The prediction problem itself was difficult. Pavement deterioration is influenced by many external factors that are difficult to observe accurately. These could be for example local weather conditions, subgrade properties, and changes in traffic composition over time. These factors introduce variability that cannot be fully captured with the available data.

Another limitation relates to the quality and consistency of the historical dataset. Some variables contained missing values and inconsistent records. These problems are partly due to the fact that some of the data entries are recorded manually, which introduces the possibility of human error.

Future research could extend the project in several directions. Additional external datasets could improve the explanatory power and generalization capability of the model for example, detailed climate and geospatial data. The uncertainty quantification methods, such as probabilistic boosting approaches, could make the predictions more useful for infrastructure management decision-making by providing confidence intervals in addition to point estimates.

From a practical perspective, the developed machine learning forecast could provide meaningful support for pavement management decisions. Although the predictions were not perfect, improved forecasting accuracy may help optimize maintenance scheduling and resource allocation across the Finnish road network. In the long term, this could reduce maintenance costs, improve road quality, and enhance traffic safety and overall transportation reliability.

7 Conclusion

The aim of this project was to develop a machine learning model for forecasting pavement rutting on the Finnish paved road network. To make the historical FTIA data suitable for this task, the original wide-format data was transformed into an event-based modelling dataset, where each observation represents the current state of a 100 m road segment and the target is a future URA measurement. This transformation made it possible to use current condition, traffic, lifecycle, lagged measurement and maintenance-related features in a supervised forecasting setting.

The final selected model was a histogram gradient boosting model trained to predict future rut depth. On the held-out test set, the model achieved an RMSE of 1.64 mm, MAE of 1.06 mm and R^2 of 0.862. These results are competitive with previous machine learning studies on rutting prediction and indicate that the event-based data transformation and feature construction were useful for modelling road deterioration.

Overall, the project demonstrates that machine learning can provide a useful forecasting tool for pavement management. The model does not replace engineering judgement or detailed maintenance planning, and its predictions remain limited by, for example, historical data quality, and missing external factors such as climate and subgrade conditions. Nevertheless, the results suggest that the developed model can support more data-driven prioritisation of road maintenance and provide a stronger basis for estimating future deterioration than simple baseline approaches.

References

[1] Krishna Singh Basnet, Jagat Kumar Shrestha, and Rabindra Nath Shrestha. "Pavement performance model for road maintenance and repair planning: a review of predictive techniques". In: *Digital Transportation and Safety* 2.4 (2023), pp. 253–267. ISSN: 2837-7842. DOI: 10.48130/dts-2023-0021. URL: http://dx.doi.org/10.48130/DTS-2023-0021.

[2] Tiago Tamagusko, Matheus Gomes Correia, and Adelino Ferreira. "Machine learning applications in road pavement management: a review, challenges and future directions". In: *Infrastructures* 9.12 (2024), p. 213. ISSN: 2412-3811. DOI: 10.3390/infrastructures9120213. URL: http://dx.doi.org/10.3390/infrastructures9120213.

[3] Seunghyun Choi and Myungsik Do. "Development of the road pavement deterioration model based on the deep learning method". In: *Electronics* 9.1 (2019), p. 3. ISSN: 2079-9292. DOI: 10.3390/electronics9010003. URL: http://dx.doi.org/10.3390/electronics9010003.

[4] Abdualmtalab Abdualaziz Ali et al. "Application of Artificial neural network technique for prediction of pavement roughness as a performance indicator". In: *Journal of King Saud University - Engineering Sciences* 36.2 (2024), pp. 128–139. ISSN: 1018-3639. DOI: 10.1016/j.jksues.2023.01.001. URL: http://dx.doi.org/10.1016/j.jksues.2023.01.001.

[5] Pedro Marcelino et al. "Machine learning approach for pavement performance prediction". In: *International Journal of Pavement Engineering* 22.3 (2019), pp. 341–354. ISSN: 1477-268X. DOI: 10.1080/10298436.2019.1609673. URL: http://dx.doi.org/10.1080/10298436.2019.1609673.

[6] Hamed Naseri et al. "Novel soft-computing approach to better predict flexible pavement roughness". In: *Transportation Research Record: Journal of the Transportation Research Board* 2677.10 (2023), pp. 246–259. ISSN: 2169-4052. DOI: 10.1177/03611981231161051. URL: http://dx.doi.org/10.1177/03611981231161051.

[7] SM Zahid Hasan. "Predictive modeling of pavement rutting using Machine Learning techniques". MA thesis. Universitat de les Illes Balears, 2024.

[8] Tamim Adnan and Abdolmajid Erfani. "Explainable AI for predicting pavement roughness under maintenance and no-maintenance scenarios". In: *Results in Engineering* 29 (2026), p. 108666. ISSN: 2590-1230. DOI: 10.1016/j.rineng.2025.108666. URL: http://dx.doi.org/10.1016/j.rineng.2025.108666.

[9] Zhen Liu, Xingyu Gu, and Wenxiu Wu. "Deterioration modeling of pavement performance in cold regions using probabilistic Machine Learning method". In: *Infrastructures* 10.8 (2025), p. 212. ISSN: 2412-3811. DOI: 10.3390/infrastructures10080212. URL: http://dx.doi.org/10.3390/infrastructures10080212.

[10] Shuting Chen et al. "Enhancing rutting depth prediction in asphalt pavements: A synergistic approach of extreme gradient boosting and snake optimization". In: *Construction and Building Materials* 421 (2024), p. 135726. ISSN: 0950-0618. DOI:

10.1016/j.conbuildmat.2024.135726. URL: http://dx.doi.org/10.1016/j.conbuildmat.2024.135726.

[11] Bing Liu et al. "A unified framework for asphalt pavement distress evaluations based on an extreme gradient boosting approach". In: *Coatings* 15.3 (2025), p. 349. ISSN: 2079-6412. DOI: 10.3390/coatings15030349. URL: http://dx.doi.org/10.3390/coatings15030349.

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

[13] Gareth James et al. *An introduction to statistical learning: with applications in R*. 2nd. Springer Texts in Statistics. New York: Springer, 2021. DOI: 10.1007/978-1-0716-1418-1.

[14] GeeksforGeeks. *ML | Gradient Boosting*. Updated May 14, 2024. GeeksforGeeks. 2024. URL: https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/ (visited on 05/22/2024).

[15] Bryan Clark and Fangfang Lee. *What is Gradient Boosting?* IBM. 2024. URL: https://www.ibm.com/think/topics/gradient-boosting (visited on 05/22/2024).

[16] Ali Juma Alnaqbi et al. "Creating rutting prediction models through Machine Learning techniques utilizing the Long-Term Pavement Performance Database". In: *Sustainability* 15.18 (2023), p. 13653. DOI: 10.3390/su151813653.

[17] Chunru Cheng et al. "Predicting rutting development using Machine Learning methods based on RIOHTrack data". In: *Applied Sciences* 14.8 (2024), p. 3177. DOI: 10.3390/app14083177.

[18] Väylävirasto. *Päällystettyjen teiden korjauksen toimintalinjat*. Väyläviraston ohjeita 10/2021. Helsinki: Väylävirasto, 2021. URL: https://aineistot.vayla.fi/api/file/ava/Julkaisut/Vaylavirasto/vo_2021-10_paallystettyjen_teiden_web.pdf (visited on 05/17/2026).

[19] Christoph Molnar. *Interpretable Machine Learning*. en. Morrisville, NC: Lulu.com, Feb. 2020.

8 Self Assessment

The final model constructed during the project departed from the initial project plan significantly due to changes made by the group. It was initially the idea to predict both the rut depth (URA) and roughness (IRI). However, after witnessing the unpredictable behaviour of IRI and discussions with the FTIA on the subject we decided to rule out IRI since the FTIA prefers that the model focuses more on URA. It was also planned that the results could be used to analyze the impact of climate change. This was sidelined due to a lack of accurate data before 2010 and to focus on building an accurate model.

The project was successful in that we were able achieve our main goal in constructing a model to predict rutting. The constructed model performs better than an average model found

in literature and discussions with the FTIA indicate that the model performs better than their previously used models. The communication between the team and the FTIA was also open and transparent. Meetings were held on regular intervals and answers to questions as well as requested documents were provided in due time.

The project was less successful in a multitude of ways. For example the auxiliary objectives were largely ignored to focus on building the model. For example the effect of increasing the weight limit of trucks on some roads was not studied and the amount of heavy traffic was used only as a feature in the model. Another thing were the slightly incoherent slides presented during the second excursion. The team had "tunnelvision" and added figures in the slides which were not properly explained leading to confusion among coursemates.

Scheduling was also a minor issue during the project and could have been done better. Most of the work done by the group centered around the deadlines and little work was done when there was no deadline in the near future. This could have been avoided with better planning and more meetings between the group. The project suffered slightly from the project manager lacking management experience.