# POLITECNICO DI TORINO

**MASTER's Degree in DATA SCIENCE AND ENGINEERING**


**MASTER's Degree Thesis**

# Predicting Road Infrastructure Deterioration with AI: A Multi-Source Data Approach

CANDIDATE
Juliana CORTÉS MENDIVIL

SUPERVISOR
Prof. Roberto GARELLO

INDUSTRIAL ADVISOR
Caterina LIA

**NOVEMBER 2025**

# ABSTRACT

The current research proposes a data-driven framework for predicting the deterioration of road infrastructure, specifically modeling the growth or spread of alligator cracking in both area and severity. This comprehensive research integrates multiple data sources including climate variables, traffic patterns, and high-resolution images from two distinct domains: the comprehensive U.S. Long-Term Pavement Performance (LTPP) dataset and a localized Italian dataset with limited historical observations.

To overcome data scarcity in the Italian domain, an instance-based transfer learning strategy was employed. A K-Nearest Neighbors (KNN) algorithm was used to locate U.S. pavement sections with environmental and traffic conditions similar to Italy. A hybrid training dataset was developed which consisted of the Italian training data augmented by the selected similar U.S. observations.

This dataset was subsequently used to train and validate three machine learning models (Random Forest, XGBoost, and LightGBM), against a Naive persistence baseline. The findings indicate that predictive effectiveness is largely determined by the severity of distress. While the Naive baseline was most effective for low-severity cracking, the Random Forest (RF) model proved to be the most accurate and robust for predicting high-severity deterioration, effectively correcting the baseline's critical tendency to underestimate structural failures.

This hybrid-data approach produces a viable and pragmatic approach for pavement management in data-scarce regions. The approach provides a more reliable forecast of high-risk segments which enables road agencies explore maintenance budgets and/or improve the management of infrastructure from reactive to preventive management.

# TABLE OF CONTENTS

# 1

# INTRODUCTION

This thesis aims to develop predictive models to analyze the deterioration of road infrastructure, focusing on the progression of alligator cracking in terms of area and severity. Using multi-source data and transfer learning from U.S. dataset (Long-Term Pavement Performance (LTPP) program [1]) to the Italian context, the study contributes to data-driven approaches for pavement management. The research was carried out in collaboration with LOKI s.r.l [2], an emerging Italian startup, within the framework of its *Asfalto Sicuro* project, which combines AI, space technologies, and vehicle-mounted sensors to automatically detect and map major pavement distresses, —including potholes, linear cracks, and alligator cracking—with the central purpose of improving road safety.

## 1.1 Problem Statement

Pavement deterioration resulting from aging, traffic loads, and environmental conditions is a persistent challenge for transportation agencies worldwide [3]. The ability to predict accurately the performance of pavements over time is required so that road infrastructure can be maintained in a broadly affordable and timely manner. Empirical models, or mechanistic-empirical models, based on historical data and expert knowledge, have long been used successfully to predict the future serviceability of pavements [4], [5]. However, these traditional empirical and mechanistic-empirical methods often struggle to capture the multi-faceted nature of the performance of pavement and its complex, nonlinear, and spatiotemporal nature under a range of loading and climatic conditions. To address these limitations, the research community has increasingly adopted machine learning (ML) and deep learning (DL) approaches for data-driven modeling of pavement deterioration [6], [7],

[8]. These models leverage diverse data sources, including historical pavement performance records, sensor measurements, traffic volumes and visual inspection data to learn predictive patterns without the need for manual rule definitions. Due to their adaptability, scalability and enhanced predictive performance, ML and DL models are already an advantageous alternative to traditional predictive approaches [9], [10].

Despite these improvements, significant challenges remain. In many countries, historical pavement performance data are either nonexistent or too sparse to support model development. Therefore, it becomes necessary to transfer knowledge from data-rich regions to data-scarce contexts [11]. However, transfer learning often yields suboptimal results due to contextual differences, particularly climatic variations that greatly affect degradation dynamics [12]. This emphasizes the need of domain-adapted approaches capable of bridging gaps between source and target environments.

Moreover, regardless of the encouraging progress, there is a clear necessity of developing machine learning models that incorporate the effects of environmental variables on the progression of certain types of pavement distresses [13]. Although numerous studies have used data as traffic, precipitation, temperature, and humidity to develop predictive models for pavement condition indices (e.g., International Roughness Index (IRI), Pavement Condition Index (PCI)), it is evident that there is more work to do in developing focus on specific types of distress that usually evolve quickly and have a much bigger impact on safety/experience.

Distresses such as potholes and alligator cracking cause deterioration of pavement structural integrity and are also immediate hazards for drivers, cyclists, and pedestrians in the system. The development of these critical distresses is always sensitive to more traditional environmental factors; especially areas that endure freeze-thaw cycles, extreme moisture, and similar conditions [14]. Consequently, machine learning models that monitor these high priority distresses will support transportation agencies in prioritizing maintenance response, improve resource allocations and earlier intervention - leading to reduced repair costs and better road usability. High-resolution environmental data combined with the distressing patterns can be integrated into a predictive model to change how agencies plan and implement pavement maintenance.

## 1.2 Objectives

### 1.2.1 General Objective

The main goal of this project is to develop and implement different predictive models for predicting the deterioration of road infrastructure, with a particular focus on the progression of alligator cracking in terms of area and severity. By integrating multi-source data—including climatic, traffic, and pavement condition information—this study seeks to develop predictive frameworks trained on a U.S. historical dataset and adapted through transfer learning strategies to the Italian context, where data availability is limited.

### 1.2.2 Specific Objectives

- **Data characterization:** To analyze and preprocess the U.S. and Italian datasets, identifying relevant variables (climate, traffic, crack area, severity) and preparing them for predictive modeling.
- **Predictive models development:** To design and train predictive models on the U.S dataset, capturing the temporal dynamics of alligator cracking deterioration under varying climatic and traffic conditions.
- **Transfer learning / fine-tuning:** To apply and compare transfer learning and fine-tuning techniques for adapting the models trained on U.S. data to the Italian dataset, enabling knowledge transfer despite limited target-domain data.
- **Evaluation of predictive performance across domains**: To assess the predictive accuracy and generalization capacity of the adapted models in forecasting crack area and severity in Italian road segments, and to benchmark them against baseline approaches.
- **Analysis of factor influence:** To analyze the relative importance of climate and traffic variables in the deterioration process, highlighting similarities and differences between U.S. and Italian conditions.
- **Validation and application:** To propose a framework for applying the models in real-world road maintenance planning evaluating their reliability for short-term forecasts in contexts with limited historical data (e.g., Italy 2025 predictions).

## 1.3 Thesis Structure

This thesis follows the next structure:

- **Literature review:** This section provides a systematic examination of the literature related to the prediction of road infrastructure degradation. The discussion is organized into two segments, the first covering a civil engineering basis by outlining the main factors affecting the performance and deterioration of asphalt concrete pavements and the second covering the application of machine learning and deep learning methods that have been developed for this purpose.
- **Methodology:** This chapter describes the datasets used in this study, as well as the preprocessing steps and feature engineering procedures applied. It also presents the methodology, detailing the predictive modeling approaches employed, including baseline models and transfer learning strategies.
- **Results and Discussion:** This section presents and analyzes the experimental results, comparing model performances across datasets and examining the influence of climatic and traffic variables on prediction accuracy. It also evaluates the transferability of models and interprets the main findings in the context of related research. Furthermore, it discusses their practical relevance for infrastructure management, acknowledges the study's limitations, and provides recommendations for future research to build on these results.
- **Conclusions:** This segment concludes the thesis by synthesizing the major findings, discussing contributions, outlining limitations, and proposing avenues of future work.

# 2

# LITERATURE REVIEW

Road networks play a fundamental role in personal mobility by enabling access to services, goods, and leisure activities, and for this reason, global economies depend on the efficient and safe operation of transportation systems [15]. Within this context, pavement represents one of the most critical components of modern transportation infrastructure, as its condition directly influences both the functionality and safety of road networks.

In Italy, the maintenance and management of the national road network is primarily entrusted to ANAS S.p.A. (Azienda Nazionale Autonoma delle Strade), a state-owned company under the Ministry of Infrastructure and Transport and part of the Ferrovie dello Stato Italiane Group. ANAS is responsible for the construction, rehabilitation, and programmed maintenance of pavements and related infrastructure across the national network, which includes state roads, expressways, and some motorways [16]. While ANAS oversees the national network, regional and municipal authorities are in charge of maintaining local and urban roads, reflecting the multi-level governance structure of road infrastructure in Italy.

At the national scale, data on road infrastructure maintenance in Italy are systematically collected and reported by the International Transport Forum of the OECD (ITF-OECD), which compiles annual statistics on transport infrastructure investment and maintenance across member countries [17]. These datasets provide a comprehensive view of Italy's overall expenditure on pavement upkeep, as they aggregate not only the resources allocated by ANAS for the national road network but also the contributions of regional and municipal authorities responsible for local and urban roads.

Recent official figures show that programmatic maintenance by ANAS exceeded €1.6 billion in 2024, while total investments and maintenance works surpassed €2.9 billion [16]. At the national level, CEIC/OECD data indicate that road infrastructure maintenance spending in 2021 was approximately €8.7 billion [17]. Despite these significant investments, deteriorated pavements remain a major safety hazard**.** Defects such as potholes, surface roughness, and alligator cracking not only accelerate structural degradation but also contribute directly to traffic accidents, loss of vehicle control, and increased risk for vulnerable users such as cyclists and motorcyclists**.** The European Road Safety Observatory states that while human factors dominate (~95%), infrastructure factors contribute to ~30% of crashes [18]. Thus, beyond the economic burden of maintenance, unsafe pavement conditions have a direct societal cost in injuries and fatalities**.** For this reason, Italian road agencies are increasingly prioritizing the monitoring of traffic, climatic, and structural factors that influence pavement performance throughout its service life, with the aim of reducing accident rates, optimizing interventions, and improving both safety and road usability.

## 2.1 Overview of Pavement Deterioration and Maintenance

A comprehensive pavement management model is vital to ensure the long-term performance, safety, and cost-effectiveness of road infrastructure. Effective management systems create a systematic approach to monitor pavement condition, forecast deterioration and prioritize maintenance interventions. These models allow decision-makers to allocate resources more effectively, lower lifecycle costs, reduce traffic disruption and improve road safety [19].

Pavement degradation over time is a natural and expected occurrence. Roads are subjected to repeated loading from daily cars, trucks, and buses, which naturally causes wear and tear. According to the AASHTO Guide for Design of Pavement Structures, the typical service life of asphalt concrete (AC) pavements under standard traffic conditions ranges from 20 to 25 years [4] .However, in practice, pavements often show signs of considerable distress before reaching their service life, in some cases, having failed within ten years (Figure 1). Such premature deterioration raises a pertinent question: what underlying factors contribute to early pavement failure?

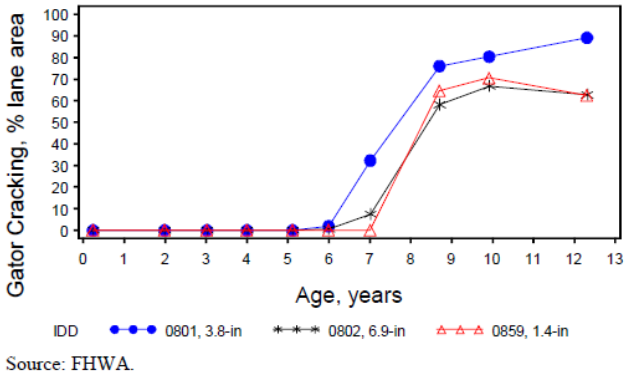



*Figure 1 Measured alligator cracking distress for SPS-8 flexible pavement projects in North Carolina [20]*

### 2.1.1 Factors Impacting Pavement Deterioration

Among the major factors affecting deterioration are traffic loading, especially from heavy vehicles, which may be a contributor to accelerated structural fatigue. The stress waves caused by repeated moving loads create permanent deformation and crack propagation that can affect the inner layers of the pavement [21]. Climatic factors, including extreme temperatures, freeze–thaw cycles, and moisture variations act to rapidly increase crack propagation through the weakening of material bonding and water infiltration into the pavement. Once cracks extend into the pavement layers, water intrusion into unbound subgrade layers promotes rapid deterioration and loss of structural integrity, ultimately leading to a reduction in the pavement's capacity to support future traffic loading [3], [20]

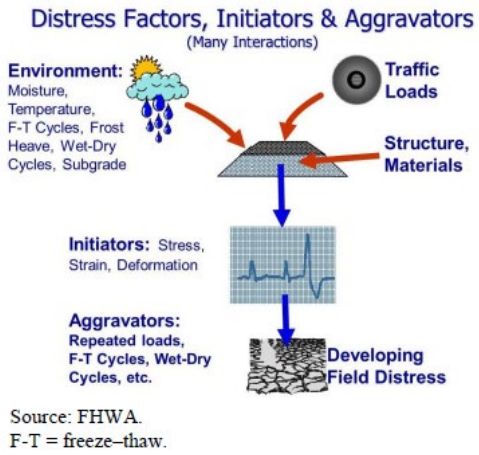



*Figure 2 Factors, initiators, and aggravators affecting pavement performance [20]*

In order to better understand these mechanisms, the U.S. Federal Highway Administration (FHWA) published an extensive investigation called "Impact of Environmental Factors on Pavement Performance in the Absence of Heavy Loads" [20]. With the basis of contributions from the Long-Term Pavement Performance (LTPP) program [1] and focused on pavement sections with very low traffic volumes (SPS-8 experiments), allowing researches to focus specifically on the effects of climate and subgrade conditions. The findings indicated that pavement deterioration results from a combination of factors, including traffic loads, construction quality, material properties, and, importantly, environmental factors [20] (Figure 2). Table 1 summarizes common asphalt concrete (AC) pavement distresses—including alligator cracking, depressions, and potholes—together with their primary and contributing causes. While, as mentioned before, it is generally accepted that traffic loads are the primary cause of the distresses, the FHWA report revealed that, environment conditions are in fact consistent contributors to the eventual damage and to the overall distress level of any given pavement [20].

*Table 1 Common asphalt concrete pavement distresses and their predominant causes [20]*

P = primary factor, C = contributing factor, N= negligible factor

| Distress | Load | Environment | | | Material | Construction |
|---|---|---|---|---|---|---|
| | | Moisture | Temperature | Subgrade | | |
| **Alligator Cracking** | P | C | C | C | C | C |
| **Bleeding** | C | C | C | N | P | C |
| **Block Cracking** | N | C | C | N | P | C |
| **Corrugation** | P | C | C | N | C | C |
| **Depression** | P | C | N | C | C | C |
| **Edge Cracking** | C | C | N | P | N | C |
| **Transverse Cracking** | N | N | P | N | C | C |
| **Longitudinal Cracking** | P | N | C | C | C | C |
| **Potholes** | C | C | C | N | P | C |
| **Raveling** | C | C | C | N | P | C |
| **Rutting** | P | C | C | C | C | C |
| **Shoving** | P | C | C | N | C | C |
| **Swelling and Bumps** | N | C | C | P | C | N |

Statistical analyses, including GLMSELECT modeling and ANOVA, highlighted the importance of freeze–thaw cycles, rainfall, moisture in the subgrade, and soil plasticity in

accelerating deterioration. The study estimated that approximately 36% of all observed damage in asphalt pavements over a 15-year period was attributable to environmental factors alone [20]. These results underscore the importance of considering environmental variables not just a contributing factor to performance, but one of the most significant driving contributors to pavement performance.

The longevity of a pavement also depends on material properties, construction quality, and maintenance practice. Pavements constructed with poor compaction, inadequate drainage, or material that does not meet minimum specifications are particularly susceptible to premature failures [22]. Likewise, delays in maintenance allow small-scale defects to evolve into major distresses like alligator cracking, rutting, or potholes and increases long-term rehabilitation costs [3].

In this context, predictive models play a crucial role. A realistic prediction model should, in principle, represent all important parameters that are known to influence the pavement performance, including traffic, climate, materials, and construction practices. Yet, due to the high complexity and nonlinear interactions among these factors, fully integrating them into a single model remains a central challenge, as evidenced by the complexities addressed in the development of the Mechanistic-Empirical Pavement Design Guide [5]. Recently, data-driven models and approaches, including statistical learning, machine learning, and deep learning, have been introduced to augment empirical and mechanistic-empirical models, as these methods have the capacity to capture multifactorial and nonlinear dynamics of deterioration [23], [24], [25]. Predictive models and related techniques allow both short-term condition prediction and long-term expected trajectories under different traffic and climate scenarios, which is indicative of the likelihood of performance [26]. Including predictive analytics into agency pavement management systems means more informed decisions for timing and type of interventions to use, which extends service life and optimizes infrastructure investments [27].

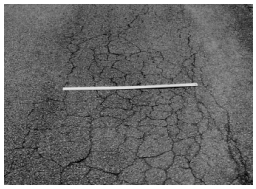
Taken together, the above evidence implies that there is a need for agencies to adopt some form of a strong pavement management framework that takes advantage of empirical knowledge and predictive modeling to anticipate deterioration and to plan maintenance actions.

### 2.1.2 Common Pavement Distresses in Asphalt Concrete Pavement

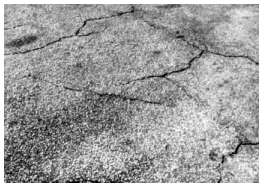
Asphalt concrete pavements incur various forms of surface and structural distresses, adversely affecting performance, safety, and ride quality, over time. According to the most recent ASTM D6433-24 standard, pavement distress refers to any external indicators of pavement deterioration resulting from the influence of traffic loading, exposure to the environment, undesirable materials and/or construction flaws [28]. In this document, a number of distinct distress types are defined (Figure 3), each with criteria for severity and extent. Some of the key distresses include:

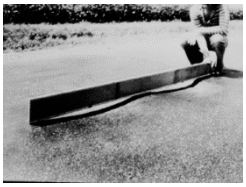
- **Fatigue (alligator) cracking:** Cracks that are interconnected and typically located in wheel paths; they result from cumulative loading that leads to asphalt layer fatigue, a classic failure mode explained by flexural theory in layered systems [22]. Over time, these cracks could develop into a network of cracks that ultimately causes localized failure or potholes.
- **Bleeding:** A surface film of asphalt or bituminous binder that rises to the surface during hot conditions, creating a slick surface. This is often a result of an unstable asphalt mix with excessive binder or low air void content [29].
- **Block cracking:** Cracks that form in a relatively large rectangular pattern that are not caused directly by car tires, but instead by shrinkage and/or temperature cycling of the asphalt; this pattern would be considered an indication of the asphalt hardening or aging over time [4].
- **Edge cracks:** Cracks located close to the edges of the pavement (generally within 0.3 to 0.5 m of the edge) that often result from weak support at the pavement edge, poor drainage, and sometimes frost action. Once the pavement edge starts to fail, raveling can occur adjacent to the crack [22].
- **Longitudinal and transverse cracks:** Cracks that are either parallel to the direction of traffic as it was paved, or cracks with the same orientation perpendicular to the direction of the traffic. These cracks can be caused by a number of factors including thermal contraction, pavement shrinkage or swelling, performing reflective cracking from underlying layers, joint between lifts or insufficient bonding, or poorly constructed joints [4].

- **Potholes and patching:** Localized surface failures when material has been lost. Failures often proceed from some sequence of fatigue cracking, or occur due to some support level failure. Potholes cause significant safety concerns. Patching refers to repair areas, but many of these areas themselves can have distress [3].
- **Rutting:** Depressions and grooves occurring in the wheel paths as a result of deformation of the asphalt, underlying base, or subgrade, all done under repeated wheel loadings. Rutting tends to be a common severity concern, because it tends to collect water and tends to be a reduction of skid resistance [3].
- **Surface defects such as raveling, shoving, bumps & swellings, corrugation:** Raveling consists of fine or coarse aggregate loss from the surface; shoving and sags are displacements caused by shear or unbalanced mix conditions; corrugation (washboarding) has regular ridges peaks and valleys normal to traffic flow. Ultimately these can will likely be caused by conditions of mix instability, insufficient compaction, insufficient support, or temperature or frost action [29].


Alligator cracking


Bleeding


Block cracking


Corrugation


Depression


Edge cracking


Transverse cracking


Longitudinal cracking


Potholes


Raveling


Rutting


Shoving


Swelling and bumps

*Figure 3 Distresses in asphalt concrete pavements [28]*

#### 2.1.2.1 Alligator Cracking

Among the numerous distresses identified in the ASTM D6433-24 standard [28], alligator cracking (sometimes known as crocodile skin cracking) is considered one of the most important signs of structural failure in asphalt concrete pavements. Unlike surface distresses, such as raveling or bleeding, alligator cracking is not merely superficial. Its presence is a classic indicator of fatigue failure, which mechanistic-empirical design principles attribute to the tensile strain at the bottom of the asphalt layer under repeated traffic loading [22], [30]. This progressive deterioration often begins with fine cracks and may quickly lead to potholes and eventually result in expensive rehabilitation if not corrected soon enough. Because of its diagnostic importance, ASTM D6433-24 outlines a detailed evaluation for alligator cracking that considers both severity and extent, a methodology aligned with major national data collection efforts like the Long-Term Pavement Performance (LTPP) program[1], [31].

This common distress is categorized into three levels of severity - low, medium, and high based on the density of the crack pattern, the width of the cracks, and the amount of surface disintegration occurring. These indicators represent an overall measure of the condition and seriousness of the potential structural failure and can be classified into three main categories: the geometric pattern, crack width, and surface condition.

- **Crack Pattern and Density**

  The progression from isolated cracks to a dense, interconnected network is the visual manifestation of the fatigue damage modeled in laboratory studies[32]

  - *Low Severity:* Hairline cracks that are widely spaced, do not link together, and are showing early fatigue, and no structural damage.
  - *Medium Severity:* A denser pattern of interconnected cracks that will make closed polygons.
  - *High Severity:* Extensive crack networks, linked, & tight, multiple crack polygons, like chicken wire, deep structural failure.

- **Crack Width**

  The width of individual cracks—from hairline cracks to wide-open cracks—provides an important sign of the severity of distress and the degree of degradation of the structure.

  - *Low Severity:* Cracks < 1/8 inch (3 mm), surface-only.

- *Medium Severity:* Cracks 1/8–1/4 inch (3–6 mm), may show minor edge wear or opening.
- *High Severity:* Cracks > 1/4 inch (6 mm), often with edge spalling, raveling, or material loss.

- **Surface deterioration**

  The condition of the pavement surface around the cracks, from intact to severely raveled, provides critical information on the stage of failure and the loss of structural integrity [29].

  - *Low severity:* Pavement surface remains smooth and intact.
  - *Medium Severity:* Slight surface roughness or flaking at crack edges.
  - *High Severity:* Severe deterioration that may contain loose materials, potholes, or deformations under load.

- **Area Extent**

  While severity captures the intensity of the cracking, extent refers to the percentage of the pavement area affected. This dimension can also give important detail to machine learning algorithms about how to weight predictions over area. Modern approaches to automated pavement management leverage image processing and deep learning to precisely quantify these exact features—pattern, width, and extent—at scale [33], [34].

## 2.2 Traditional Predictive Models for Pavement Performance

Models for predicting pavement performance are essential in pavement engineering, as they allow engineers to make predictions about pavement behavior over time under defined traffic and environmental conditions. Three primary categories of traditional predictive models have been developed: empirical models, mechanistic models, and mechanistic-empirical (M-E) models, with the latter two being most widely used.

Empirical models, historically the foundation of pavement design, are primarily derived from the analysis of extensive field and experimental data, establishing statistical relationships between observed pavement performance and influential design variables such as traffic loading, material properties, and climatic factors [22]. A notable example is the AASHTO 1993 Pavement Design Guide, based on the results of the AASHO Road Test, which links key variables—such as traffic (often expressed as cumulative equivalent single

axle loads, ESALs), subgrade support, and desired reliability—to pavement thickness and performance predictions [4]. The World Bank's HDM-4 system is another example, which also relies heavily on empirical deterioration models that agencies often calibrate to local conditions [35]. Although empirical models are straightforward and data-efficient, they lose accuracy when applied to conditions differing from those for which they were developed. [4], [22]

The mechanistic–empirical (M–E) approach was created to overcome the limitations associated with purely empirical methods by integrating them with engineering mechanics. M–E models simulate the three-dimensional responses of pavements to traffic and climate (i.e., stresses, strains, and deflections) and link the responses to actual field performance through transfer functions [5].The AASHTO Mechanistic-Empirical Pavement Design Guide (MEPDG) is a premier example that incorporates detailed traffic, climate, and materials inputs into pavement performance predictions [5]. Despite being more adaptable and scientifically sound than their purely empirical predecessors, M-E models require extensive high-quality data and localized calibration to ensure accurate predictions [36].

## 2.3 Applications of Machine Learning in Pavement Evaluation

Employing Machine Learning (ML) and Deep Learning (DL) have significantly advanced flexible pavement evaluation by uncovering complex, non-linear relationships in large datasets from non-destructive testing (NDT) and historical performance records [9]. Unlike traditional statistical approaches, these models characterize complex patterns influencing pavement deterioration and overall performance over time.

ML models such as Random Forest (RF), Gradient Boosting (e.g., XGBoost, LightGBM), and Support Vector Machines (SVM) have been widely applied to predict pavement condition indicators like the International Roughness Index (IRI), Pavement Condition Index (PCI), cracking, rutting, and faulting [9], [10], [24]. Using databases such as the U.S. Long-Term Pavement Performance (LTPP) [1], these models have demonstrated strong accuracy and interpretability [23].

RF models, for instance, have illustrated the ability to learn nonlinear interactions and identify important predictors such as annual average daily truck traffic (AADTT), temperature, and thickness of the layers across multiple distress mechanisms [9]. A recent

study of continuously reinforced concrete pavements presented a random forest framework for predicting multi-distress [37]. Results showed that RF successfully models cracking, faulting, and roughness distress at the same time, generated interpretable rankings of variable importance for different, transferable to flexible pavements [37].

Likewise, Gradient Boosting techniques like XGBoost and LightGBM have displayed greater predictive accuracy through an iterative process of correcting the previous model's mistakes [38], [39]. In PCI prediction studies, boosting algorithms have continually outperformed linear and single tree baselines as well. One example is the FCM-XGBoost model in predicting PCI, which applied a fuzzy c-means clustering method in which the pavement sections were grouped into homogenous clusters before training the XGBoost model, resulting in better predictive robustness and interpretability than traditional methods [40]

These ensemble and hybrid machine learning methods are able to achieve not only high predictive performance, but also provide insight into the relative impacts of design, environmental, and traffic attributes, providing useful feedback for pavement management and design validation.

## 2.4 Applications of Deep Learning in Pavement Evaluation

Deep learning techniques have remarkably improved flexible pavement assessment by analyzing complex spatial and temporal dependence across a variety of data types. Deep Neural Networks (DNNs) have been utilized to forecast key condition metrics such as the Pavement Condition Index (PCI) and International Roughness Index (IRI), using heterogeneous data from pavement management systems, including traffic, climate, materials, and maintenance data. Boonsiripant et al. [41] demonstrated that DNNs can achieve comparable accuracy to graph convolutional models for IRI prediction, and Radwan et al. [42] demonstrated improved prediction for PCI over traditional regression methods.

Temporal models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are effective for predicting deterioration by learning sequential dependencies in historical condition data. For example, it was developed a robust, interpretable, and high-accuracy LSTM + Multi-Head Attention framework [26] that outperforms traditional machine learning and standard deep learning models in predicting pavement IRI.

Similarly, convolutional and object-detection architectures—such as CNNs, U-Net, Mask R-CNN, and YOLO—have revolutionized visual inspections of pavements through automated crack detection, classification, and quantification with excellent accuracy based on pavement imagery [33], [43].

## 2.5 Transfer Learning in Pavement Performance Prediction

Conventional ML models, including sophisticated deep learning architectures, require substantial quality historical datasets for training to achieve acceptable accuracy and generalizability. This requirement clearly limits their feasibility in several real-world contexts, particularly for developing countries or newly developed road networks where historical datasets are scarce or non-existent [11], [12].

To mitigate this issue, Transfer Learning (TL) is a contributing paradigm that enables knowledge to be transferred from a data-rich (source) to a data-scarce (target) domain. The fundamental concept is to leverage the knowledge and experience gained solving one problem (the source task, typically with abundant data) to a problem that is different but related (the target task, typically with limited data) [44]. This approach effectively transforms traditional " learning from scratch" to performing cumulative learning, thereby reducing the need for large target-domain datasets [12].

A key methodology in this area is instance-based transfer learning. Algorithms such as TrAdaBoost.$R^2$ [45] and its enhanced version, the Two-Stage TrAdaBoost.$R^2$ [46] that operates by iteratively re-weighting data from the source domain during training, assigning higher weights to source instances that are similar to the target data while reducing the weight of dissimilar ones. Successful applications have demonstrated the high potential of this approach, such as using the extensive U.S. Long-Term Pavement Performance (LTPP) database [1] to accurately predict the International Roughness Index (IRI) for highways in China [12] and Portugal [11], in both cases significantly outperforming models trained on local data alone.

However, transferring pavement performance models across different geographic contexts (e.g., from the U.S. to Italy) faces several challenges. First, climatic differences that include variations in temperature and precipitation drive unique mechanism of deterioration that may not be fully captured by models trained on source data. Second, differences in

pavement design standards, materials, and construction methods create a fundamental mismatch in how input features relate to performance outcomes. Lastly, these climatic and design variations result in a data distribution shift, where the key variables, such as IRI values and traffic loads, may no longer have the same statistical properties in the source and target domains leading to reduced model generalizability if adaptation is not properly handled.

Recent reviews emphasize the importance of TL and domain adaptation techniques for improving pavement performance modeling [11], while newer methods such as ISTRBoost propose advanced re-weighting strategies to further mitigate negative transfer effects [47]. These developments underscore TL's growing importance as a feasible approach to address data scarcity in pavement performance prediction.

## 2.6 Data sources in Pavement Performance Modeling

Reliable predictions of pavement deterioration require complete and systematically organized datasets, and predicting alligator cracking progress in terms of initiation and deterioration will depend on climate, traffic load, material properties, and construction practices [48]. Because the process of alligator cracking initiation and deterioration occurs over time, predictive models need to incorporate and account for these factors over multiple years and within local contexts. It is also essential to collect repeated observations on the same road segments, as even roads within the same area are likely to experience different traffic compositional loads and subsequent aging behaviors[49]. However, temporally consistent and regularly updated datasets remain scarce despite the existence of large-scale databases such as the Long-Term Pavement Performance (LTPP) program [1]. Recently, private initiatives like LOKI [2]have begun addressing this gap by developing georeferenced databases that document distress types, images, and affected surface areas in regions such as Piedmont, Italy.

# 3

# METHODOLOGY

This study follows a methodological framework that takes a multi-source, data-driven approach to model and predict the progression of alligator cracking in terms of area and severity on asphalt pavements. The approach integrates heterogeneous datasets from two geographic domains - United States (source domain) and Italy (target domain) - and applies machine learning and transfer learning techniques to account for contextual differences.

## 3.1 Overall Methodological Framework

The methodological pipeline proceeds through five key stages:

- **Data acquisition and integration**

  The study combines pavement performance data, climatic variables, and traffic information from two distinct sources: the Long-Term Pavement Performance (LTPP) database for the United States [1] and a proprietary Italian dataset developed by LOKI s.r.l. under the *Asfalto Sicuro* project [2].

- **Exploratory data analysis**

  Because the data originate from different formats and collection methodologies, some preprocessing was needed to ensure a consistent structural framework. Duplicated values were eliminated, missing values imputed and the logical process of deterioration was revised. Then, the data structure was inspected, using histograms and violin plots to evaluate the shape, skewness and the outliers of the variables; correlation heatmaps to identify potential dependencies among variables,

- **Comparison between datasets**

  This section gives a side-by-side examination of the U.S. (LTPP) and Italian datasets in terms of differences in scale, degradation states, environmental and traffic characteristics, and statistical relationships, to help gain insights into how the characteristics of data influences our modeling results.

- **Experimental setup and modeling framework**

  In this chapter, the experimental design and modeling framework utilized in the thesis is described. It discusses how data is organized into subseries, how training/testing sets are organized, and how cross-domain standardization and KNN-based similarity mapping bring the LTPP context to Italy. It describes the supervised one-step, multivariate regression task, elaborates on methodological obstacles (limited target data, domain shift and inter-severity dependence) and model selection criteria. The chapter ends with the candidate models (RF, XGBoost, LightGBM, KNN), baseline Naive approach, multi-output approach, and evaluation protocol ($R^2$, RMSE, MAE) and overfitting checks ($\Delta R^2$) being presented.

- **General methodological limitations**

  In this part, it is discussed the main methodological constraints encountered during the labeling of alligator cracking severity. It highlights the practical difficulties of applying standardized assessment criteria—such as subjectivity in expert judgment, image redundancy, perspective distortions, lack of georeferencing, and variability in pavement structures—that collectively limit the consistency and reproducibility of severity classification.

## 3.2 Data acquisition and integration

### 3.2.1 United States (LTPP) Dataset

The Long-Term Pavement Performance (LTPP) program is one of the most comprehensive pavement monitoring initiatives ever conducted. It was established by the Federal Highway Administration (FHWA) in cooperation with the American Association of State Highway and Transportation Officials (AASHTO) to improve understanding of pavement behavior under varying environmental, structural, and traffic conditions across North America [1].

The LTPP program systematically collects, processes, and publishes long-term data from more than 2,500 pavement test sections distributed across the United States and Canada. Each section is monitored continuously to record variables concerning related to:

- Pavement structure, including layer thicknesses and material properties.
- Traffic loading (usually denoted by Average Annual Daily Truck Traffic, or AADTT).
- Climatic conditions (precipitation, temperature, and freeze-thaw cycles, among others).
- Surface distress, covering types and severities of cracks (including alligator cracking, block cracking, rutting, and potholes).
- Roughness and deflection measurements, through indices like the International Roughness Index (IRI) and Falling Weight Deflectometer (FWD) tests.

This large amount of data is valuable because it enables the development of models that can derive relationships between the parameters collected and the evolution of the network condition. For this research, the LTPP database was employed as the source domain for training and calibration of machine learning models. Specifically, pavement sections containing records of alligator cracking between the years 1980 - 2021 were extracted as this distress type reflects the progression of load-associated fatigue failure in asphalt layers. The selected data include:

- Pavement distress variables: area of alligator cracking classified into low, medium, and high severity ($m^2$).
- Environmental variables: mean annual temperature (°C), total annual precipitation (mm).
- Traffic variables: (AADTT) annual average daily truck traffic (trucks/day).
- Temporal variable: survey year.

### 3.2.2 Italy – San Sebastiano da Po, Piemonte

The target domain corresponds to and Italian dataset focused on the municipality of San Sebastiano da Po, located in the Piemonte region. It was developed using data from the *Asfalto Sicuro* project provided by LOKI S.r.l.[2], combined with regional climatic and traffic information from Piemonte [50], [51].

### 3.2.2.1 Visual Inspection Data

The *Asfalto Sicuro* system, created by LOKI S.r.l.[2], automates the manual visual inspection process to detect pavement defects with an AI-based system. The inspection is performed using a plug-and-play hardware system mounted on a regular vehicle, which employs high-resolution cameras and IMU sensors, as well as authenticated GNSS (Galileo HAS/OSNMA) technology, to collect data. This allows for the acquisition of geolocated images at normal driving speeds, up to 90 km/h, without causing traffic delays. The images that were acquired are processed with deep learning algorithms to detect and classify road damages. The system distinguishes:

- Potholes (low, medium, high severity, determined by depth and diameter thresholds;)
- Cracks (linear)
- Alligator cracks (fatigue-related patterns on asphalt)
- Other anomalies (manholes, road markings, architectural barriers.)

When identified and localized automatically, each defect has coordinates confirmed by the GNSS module and stored with the image frame containing the area in cm². In the Figure 4 it is shown an example of the *Asfalto Sicuro* system interface displaying a Multi Crack defect on Via Rigonda with its localization, surface area, and photographic evidence.

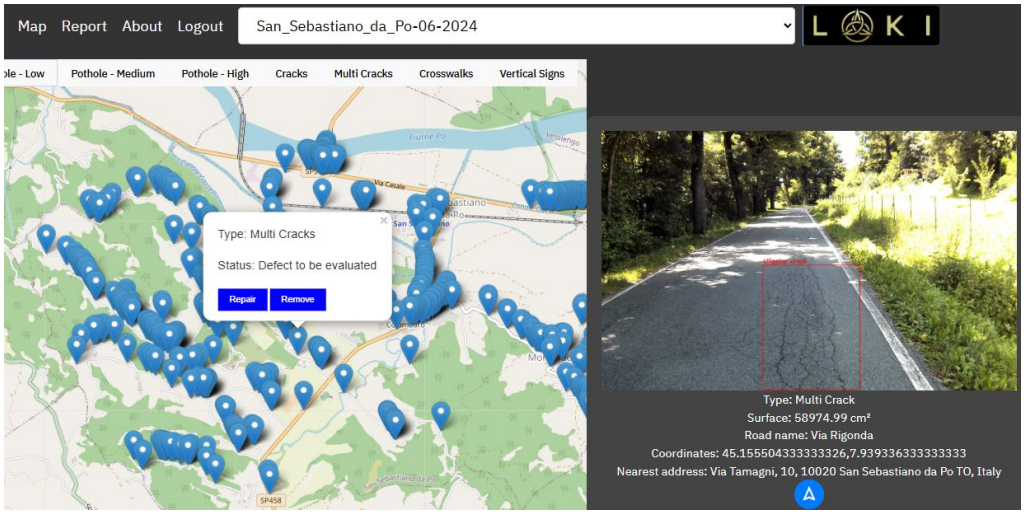



*Figure 4 Example of Asfalto Sicuro system interface [2]*

### 3.2.2.2 Labeling Methodology and Severity Classification

As mentioned before, the alligator cracking defect was selected as the focus of analysis in this study. Although this defect is included in the *Asfalto Sicuro* project, unlike the potholes, it does not currently have a defined severity rating classification system. To address this, the pavement sections that were identified with multicrack campus failures, were sampled, and the images of the sections collected.

Using Roboflow [52] as instance segmentation and dataset management platform, along with the engineering experience and subject matter expertise, the images were manually outlined and based on the distress severity, a low, medium or high severity level label was assigned to that area (see Figure 5). The outcome of this feature is a structured dataset with the areas of the alligator cracking by severity that can now be used to enable machine learning–based predictive modeling for pavement distress.

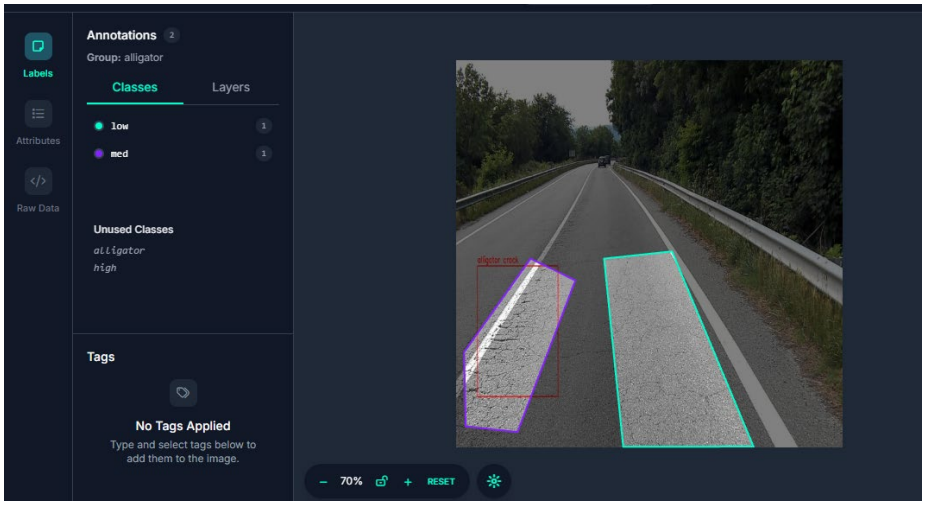



*Figure 5 Example of annotations in Roboflow.*

### 3.2.2.3 Environmental and Climate Data

The environmental variables applied in this study - total annual precipitation and mean annual temperature - were obtained from the ARPA Piemonte meteorological database (Agenzia Regionale per la Protezione Ambientale del Piemonte) through the MeteoWeb interactive map service [50] which provides long-term time series data from the regional weather station network. For the area under study, the Castagneto Po meteorological station is selected as the nearest and most representative source of climatic data. Annual aggregated values are

extracted for the period 2002–2024, and since no data are available for 2025, this value is extrapolated based on the historical trend observed over the 2002–2024 series.

This process ensures that all climatic inputs provided to the predictive model during analysis are uniform, spatially representative, and aligned with the temporal resolution values in the *Asfalto Sicuro* dataset.

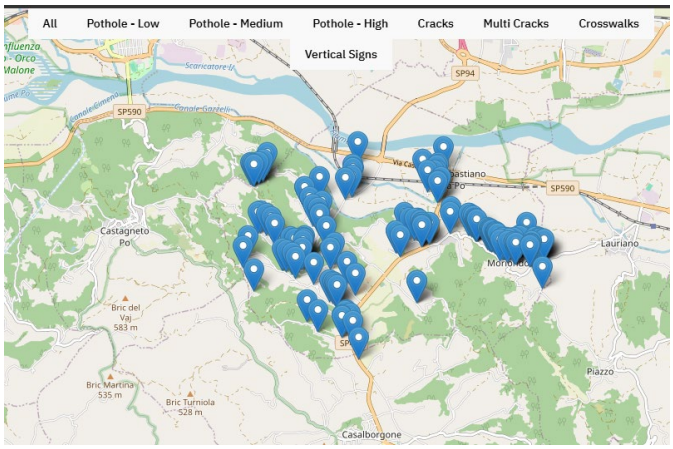



*Figure 6 2024 Multicracks map given by Asfalto Sicuro [2]*

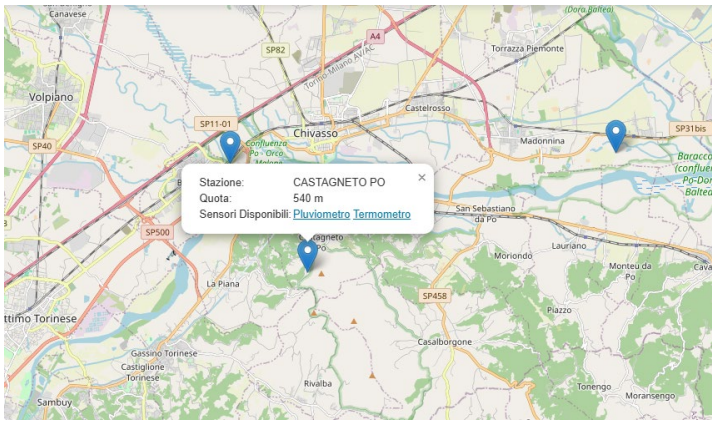



*Figure 7 The closest meteorological station to the segments [50]*

### 3.2.2.4 Traffic Data

The traffic variable, represented by AADTT (Average Annual Daily Truck Traffic), is obtained from the Geoportale Piemonte database, which contains open-access geospatial datasets on regional mobility and infrastructure. In particular, the variable is drawn from the dataset entitled "Traffico Giornaliero Medio" through the GeoNetwork portal of Regione Piemonte [51]. The dataset contains georeferenced measurements of traffic intensity across

the regional road network, comprising the average daily flow of light vehicles, the average daily flow of heavy vehicles, and the total average daily flow of both. The data is downloaded in GeoPackage (GPKG) format (see Figure 8) and subsequently processed using QGIS to convert and organize them into a tabular format suitable for integration with the road distress dataset.

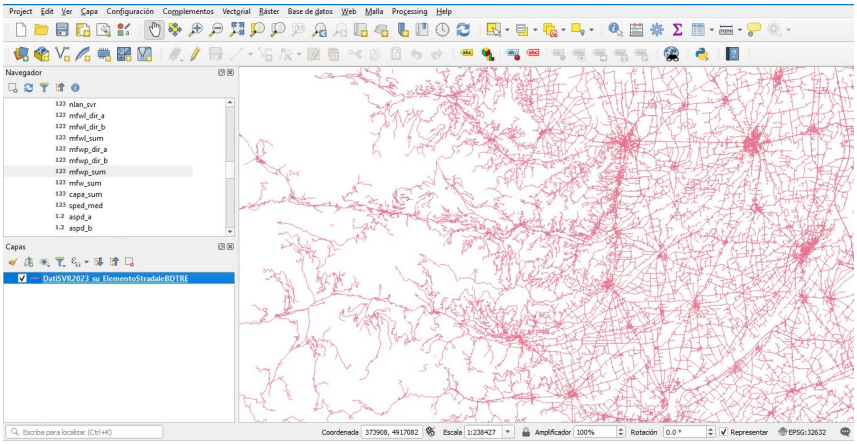



*Figure 8 GeoPackage data from QGIS*

The chosen traffic variable is the average daily traffic flow of heavy vehicles. It is an important parameter in the pavement performance evaluation since heavy trucks produce considerably greater axle loads and cyclical stress on the pavement structure than light vehicles. These axle loads result in quicker fatigue damage, particularly in flexible pavements, and are a primary contributor to load-associated distresses such as alligator cracking and rutting [4], [22]. Thus, the AADTT is a representative for total mechanical demand applied to the pavement surface, which allows the model to include the direct effects of traffic loading on the rates of distress.

The available records cover the period 2015–2023; therefore, a linear extrapolation was used to predict the values for 2024 and 2025 ensuring temporal alignment with the *Asfalto Sicuro* observations. Since not all road sections included direct traffic measurements, a further feasibility check was undertaken through Google Maps, where the road hierarchy (main, secondary, local) was identified to assign the most representative traffic flow to each segment (see Figure 9), maintaining spatial consistency of AADTT across the dataset. The extrapolated results in Table 2 represent the projected AADTT values for the analyzed road sections, extending the historical trend of truck flow in the study area.

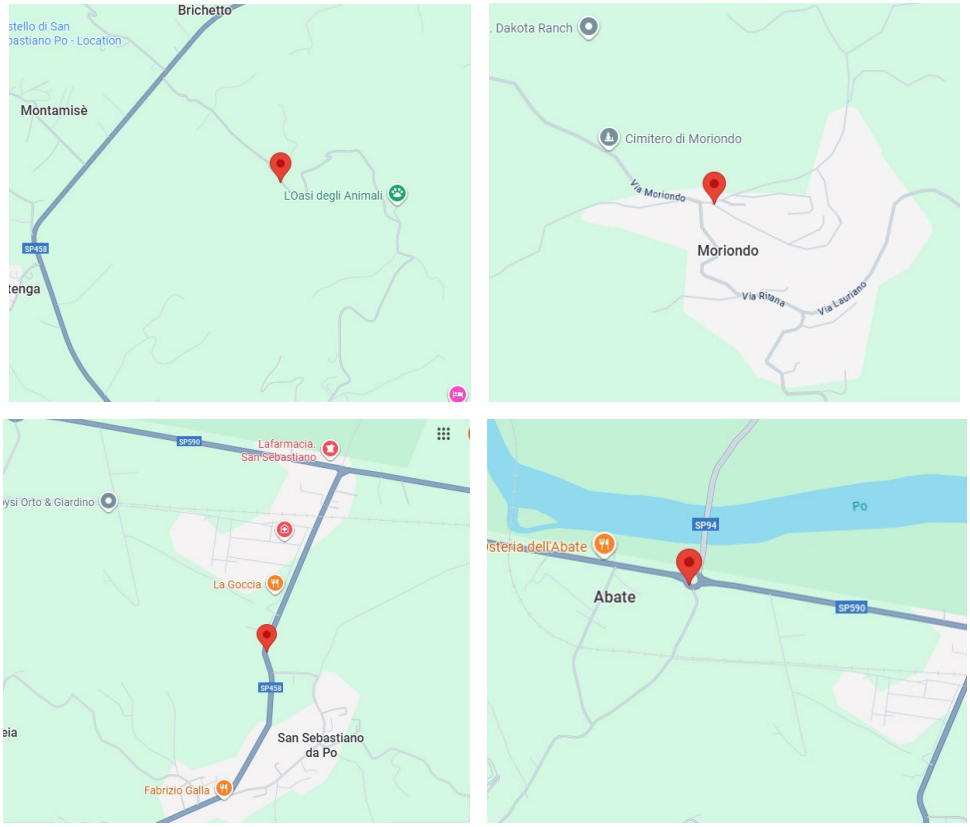



*Figure 9 The road hierarchy identified by using Google Maps*

*Table 2 AADTT given by Geoportale and predictions for 2024 and 2025 years*

| Type | Street name | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 |
|---|---|---|---|---|---|---|---|---|---|
| **1** | Ricca | 10 | 12 | 19 | 11 | 3 | 8 | 13 | 18 |
| **2** | Bertola | 20 | 24 | 30 | 11 | 6 | 20 | 34 | 48 |
| **3** | Chivasso | 210 | 233 | 721 | 713 | 568 | 737 | 906 | 1075 |
| **4** | Casale | 930 | 948 | 1634 | 1577 | 1515 | 1397 | 1316 | 1310 |

## 3.3 Exploratory Data Analysis

### 3.3.1 Data Preprocessing

Prior to model development and data analysis, a thorough data preprocessing step was performed to check for internal consistency and reliability of the dataset. First, duplicate records were removed and missing values were eliminated to avoid inconsistencies in values. In the case of Italian dataset there weren't missing values, on the contrast, for the LTPP dataset, for the distress variables, the number of missing values was very low, comprised of

six and seven missing values for low and medium severity, and five for high severity, and hence, those records were eliminated without compromising the overall representativeness of the dataset. Conversely, the traffic variable (AADTT) contained 363 records for zero values, which are implausible given that they represent average daily truck traffic. As a result, the 363 zero values were addressed by applying an interpolation procedure, using the sequential temporal values separately for each road segment, allowing the ability to substitute within realistic estimates that maintained the temporal continuity of the series.

Considering the time series nature of the data, where each record corresponds to a specific year of observation, it was also necessary to verify the logical evolution of pavement deterioration. The total area of cracking (calculated as the sum of low, medium, and high severity) was examined to confirm that it was non-decreasing over time within each subseries (indicating a physically plausible progression of damage). In addition, transitions from one severity level to another were explored to check if the patterns were markedly reasonable: when low severity area decreased, the medium severity area increased; when the medium severity area decreased, the high severity area increased. Verifying these logical process checks was critical in assuring that the data represented plausible processes of deterioration without anomalies that would threaten inference in later modeling tasks.

### 3.3.2 Descriptive statistics

This phase includes several descriptive statistics, which outlines the characters of tendencies, variability, and the distribution of major variables of interest including traffic intensity (AADTT), total annual precipitation, mean annual temperature, and distressed areas by severity level (low, medium, high). Through this analysis, potential outliers, data inconsistencies, and underlying trends were identified, supporting the understanding of the physical and environmental factors influencing pavement deterioration. The descriptive analysis includes graphical representation tools such as histograms, boxplots, scatterplots, and time-series visualizations that supplemented the numerical analyses and prioritization of observations of interest for cross comparison of variable relationships among the U.S. and Italian datasets.

### 3.3.2.1 United States (LTPP) Dataset

- **Statistical Summary**

The descriptive statistics of the variables from LTPP dataset are shown in the Table 3, which includes a total of 11,993 annual observations of pavement sections across 49 U.S. states, spanning the years 1980 to 2021. The variable construction number (cons_num) takes on values between 1 and 14, indicating multiple maintenance, or reconstruction events for some sections. However, the median is 2, suggesting that moderate number of segments experienced only one or two interventions during the monitoring period.

*Table 3 Descriptive statistics of LTPP dataset.*

| | state | year | cons_ num | area _low | area_ med | area_high | precip | temp | AADTT |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 11993 | 11993 | 11993 | 11993 | 11993 | 11993 | 11993 | 11993 | 11993 |
| **mean** | 24.89 | 2000.18 | 2.47 | 17.34 | 9.63 | 7.24 | 900.43 | 14.47 | 929.76 |
| **std** | 15.67 | 7.41 | 1.49 | 39.85 | 35.06 | 44.7 | 437.11 | 4.9 | 1071.97 |
| **min** | 1 | 1980 | 1 | 0 | 0 | 0 | 59.06 | -2.18 | 2.33 |
| **25%** | 10 | 1994 | 2 | 0.12 | 0 | 0 | 474.67 | 10.44 | 284 |
| **50%** | 26 | 1999 | 2 | 2.4 | 0 | 0 | 1011.05 | 14.53 | 531 |
| **75%** | 39 | 2005 | 3 | 15.5 | 0.9 | 0 | 1231.4 | 18.63 | 1170 |
| **max** | 49 | 2021 | 14 | 564.3 | 471.1 | 816.6 | 1824.82 | 24.26 | 10234 |

For the deterioration indicators, the mean area of alligator cracking with low, medium, and high severity (area_low, area_med, area_high) is 17.34 $m^2$, 9.63 $m^2$, and 7.24 $m^2$, respectively, with relatively large standard deviations—especially for high severity (44.70 $m^2$). The 75th percentile indicates that in most instances the observations are low, with high-severity cracking absent in a large portion of records (median = 0).

In terms of environmental conditions, the mean annual temperature averages 14.47 °C ranging from −2.18 °C to 24.26 °C and in terms of total annual precipitation it averages around 900 mm demonstrating the climatic diversity of the monitored sites. The traffic variable (AADTT) has a mean of 930 heavy vehicles per day a standard deviation of 1071 − indicating a large amount of dispersion in traffic loading across the network. Overall, the data set has a wide range of climate and operational conditions.

- **Univariate distributions**

Figure 10 illustrates the frequency distribution or histogram of the target variables including alligator cracking area for low, medium, and high severity level in the LTPP dataset. All three histograms are clearly right-skewed having a very large portion of sections with small values of cracking area, and only a few sections with very large values of cracking area (over 200 $m^2$).

The plots for low and medium severity level are very tight and concentrated around zero, representing both the reasonably good condition and the limited historical timeframe (2024-2025). The high severity level distribution is right-skewed as well, but exhibits a slightly longer tail which indicates that there are some serious levels of distresses on a few sections.

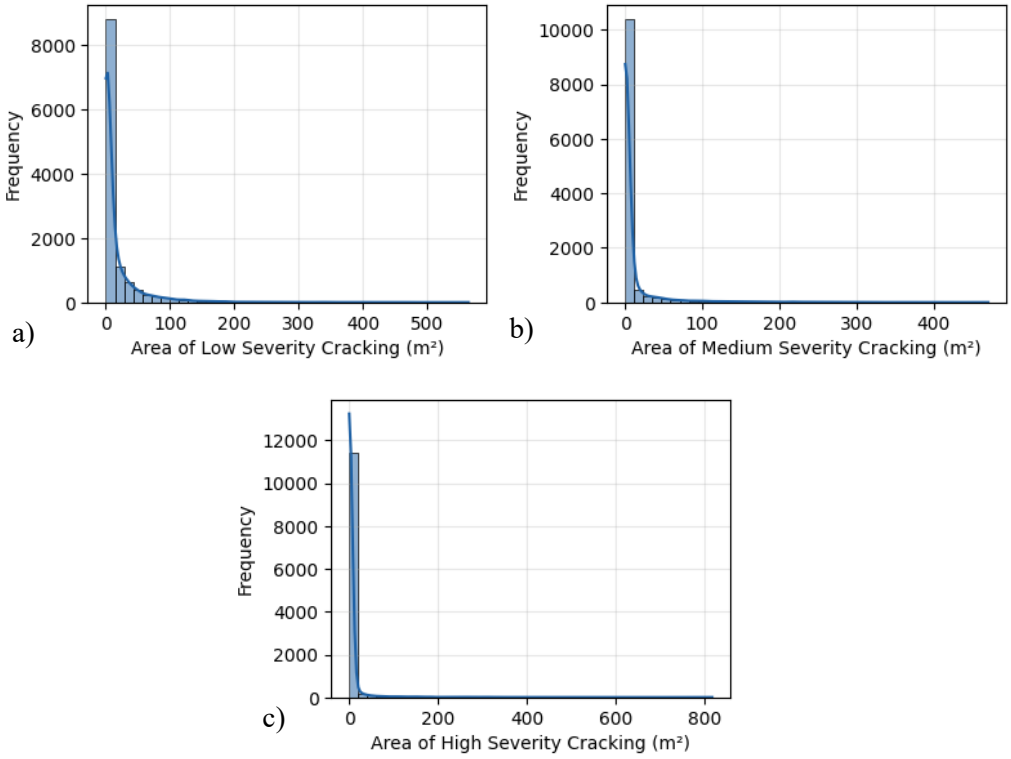



*Figure 10 Distribution of the Alligator Cracking areas for LTPP Dataset a) Low severity b) Medium severity c) High severity*

- **Violin Plots**

As shown in Figure 11, the violin plots shows that the low-severity group has the largest variability, with a wider distribution of values and a few extreme observations. This indicates early-stage cracking is the most prevalent and variable form of deterioration. Compared to low-severity, the medium- and high-severity violins are much narrower, indicating advanced cracking is less frequent and occurs within a narrower distribution of values.

In the case of total annual precipitation, mean annual temperature, and Average Annual Daily Truck Traffic (AADTT), the violin plots in Figure 12 illustrate that the precipitation values are mainly concentrated between 500 and 1500 mm, with a relatively symmetric distribution centered around 1000 mm that means the rainfall is consistent for most of the road sections. The annual average temperature followed a bell-shaped distribution, with most data points between 5 °C and 20 °C and a median around 15 °C, signifying a predominantly mild climate. The AADTT distribution on the other hand exhibits a strong right-skewed distribution with most values below 2000 trucks per day, while the highest values exceeded 9000 trucks per day, meaning there is a large variability in traffic intensity.

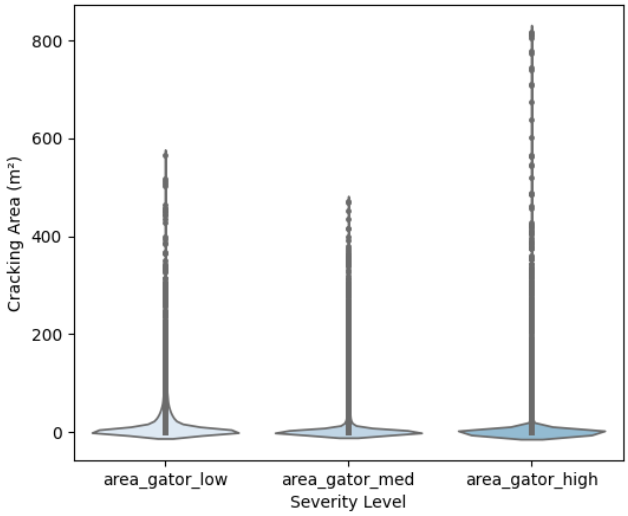



*Figure 11 Violin plot of alligator cracking areas for LTPP Dataset*

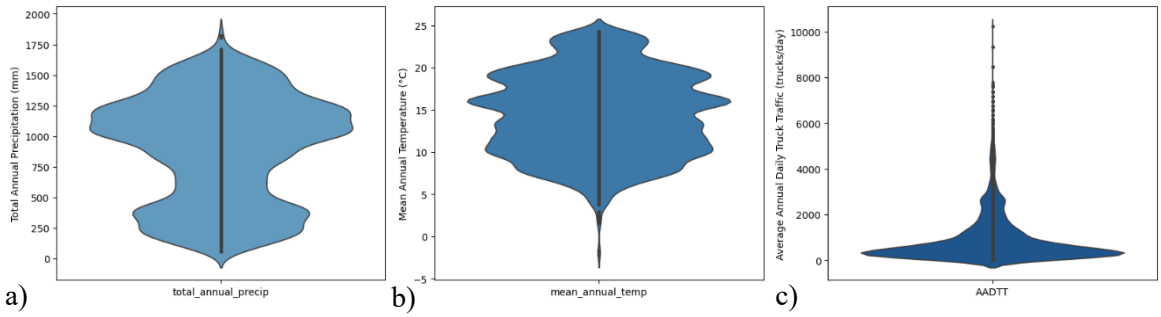



*Figure 12 Violin plots for LTPP Dataset a) Total annual precipitation b) Mean Annual Temperature c) AADTT*

- **Bivariate analysis**

The correlation heatmap for the LTPP dataset (U.S.) in Figure 13, evidences a generally weak linear relationship with the examined variables. The three severity levels of alligator cracking

(low, medium, and high) show very weak correlations between each other, with their correlation coefficients very close to zero.

AADTT (Average Annual Daily Truck Traffic) exhibits a weakly positive, and minimal correlation with high severity cracking (0.09). The data suggests that there is little relationship between the traffic class intensity and the deterioration of the pavement at the more severe levels observed. Environmental variables such as total annual precipitation and average annual temperature also showed weak or negligible correlations with the distress variables.

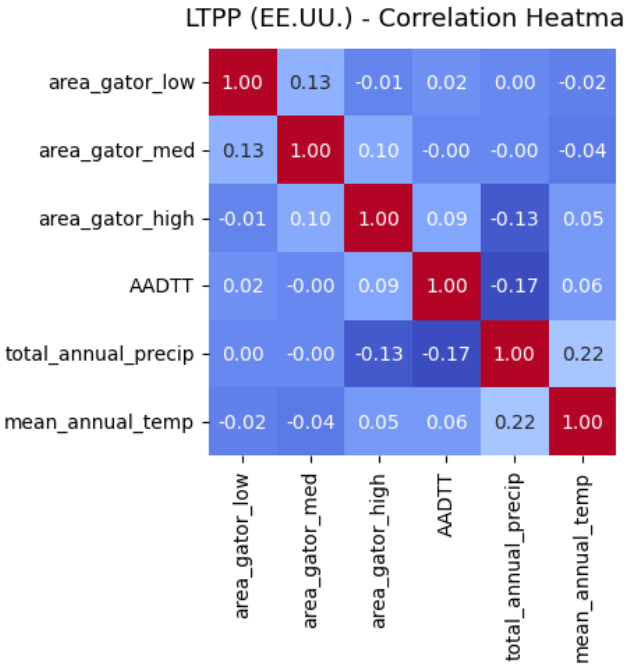



*Figure 13 Correlation heatmap of LTPP dataset*

### 3.3.2.2 Italian Dataset

- **Statistical Summary**

The descriptive statistics of the Italian dataset shown in the Table 4, contains 170 observations corresponding to pavement sections monitored during the years 2024 and 2025. All records come from San Sebastiano da Po and have the same construction number equal to 1, indicating that all sections belong to the same maintenance cycle and no rehabilitation event was recorded for these sections.

*Table 4 Descriptive statistics of Italian dataset.*

| | state | year | cons_ num | area _low | area_ med | area_high | precip | temp | AADTT |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 170 | 170 | 170 | 170 | 170 | 169 | 170 | 170 | 170 |
| **mean** | 100 | 2024.5 | 1 | 0.33 | 1.18 | 1.92 | 1173.42 | 13.52 | 361.74 |
| **std** | 0 | 0.5 | 0 | 0.95 | 1.8 | 1.75 | 62.86 | 0.25 | 548.67 |
| **min** | 100 | 2024 | 1 | 0 | 0 | 0 | 1110.75 | 13.28 | 13 |
| **25%** | 100 | 2024 | 1 | 0 | 0 | 0 | 1110.75 | 13.28 | 34 |
| **50%** | 100 | 2024.5 | 1 | 0 | 0 | 1.71 | 1173.42 | 13.52 | 48 |
| **75%** | 100 | 2025 | 1 | 0 | 1.71 | 2.98 | 1236.1 | 13.77 | 906 |
| **max** | 100 | 2025 | 1 | 5.97 | 8.59 | 7.89 | 1236.1 | 13.77 | 1316 |

Regarding pavement deterioration, the average areas of cracking indicate that low-severity cracking is nearly absent (mean = 0.33 m²), while medium and high severities have slightly elevated average areas (1.18 m² and 1.92 m², respectively), implying that most of the surveyed sections present moderate to advanced stages of deterioration. However, the high standard deviations (especially for medium and high severities) indicate a high level of variability for segments, even increasing cracking areas up to 8.6 m² in some cases.

Environmental variables display relatively stable conditions, including an annual average precipitation of approximately 1173 mm and a mean temperature of 13.5 °C, both with minimal dispersion, indicating a consistent climatic regime across the analyzed network. The traffic variable (AADTT) displays high variability (mean = 362; std = 549), ranging from very low to more than 1300 heavy vehicles per day, and it may help explain some of the heterogeneity in the observed levels of deterioration.

- **Univariate distributions**

Figure 14 shows the distribution of alligator cracking areas for Italian dataset low, medium, and high-severity levels. Each distribution shows a significant right skew, indicating a predominance of pavement sections having a small area of cracking and a much fewer number of pavement sections showing significant cracking area. For the low-severity cracking, values aggregated near to zero, which indicates that the deterioration in the pavement is mostly early-stage. For medium-severity, the alligator cracking distribution shows wider variability, which provides an indication that there is a growing area of surface area damaged as alligator cracking develops. Lastly, the high-severity cracking produces the

widest distributed values, including a few examples that exceeded 6 m², which would imply that more advanced structural damage has occurred in these pavements.

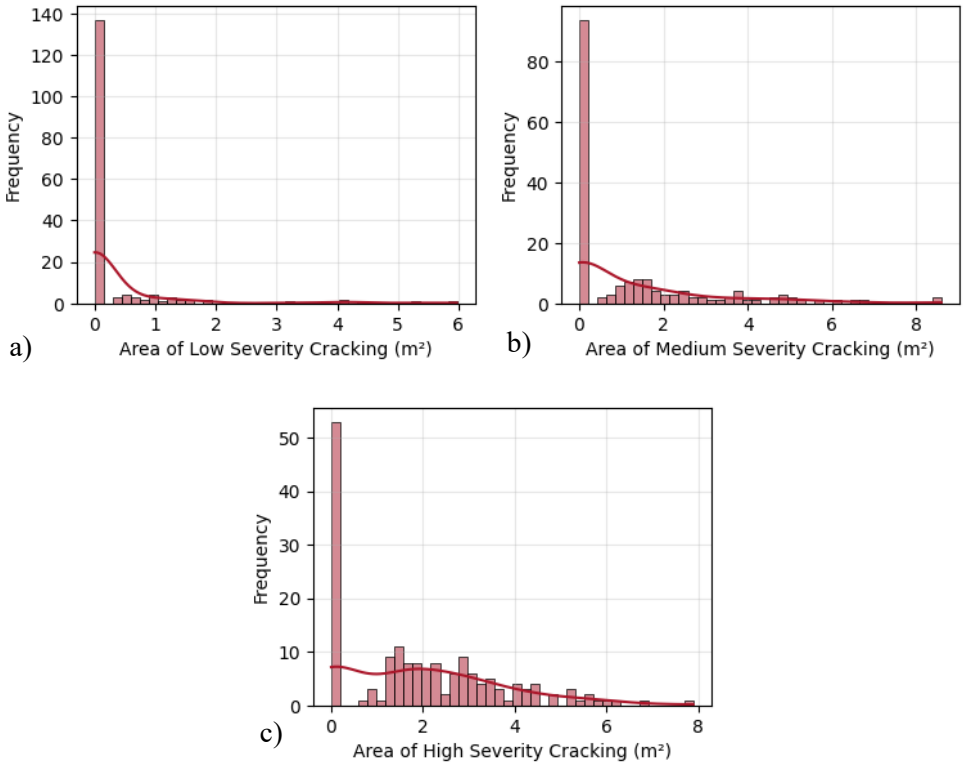



*Figure 14 Distribution of the Alligator Cracking areas for Italian Dataset a) Low severity b) Medium severity c) High severity*

- **Violin Plots**

The distributions illustrated in Figure 15 show a fairly continuous behave across three severities, with the majority of the observations concentrated at small areas of cracking, but still having noticeable density progressively extending towards the higher values. Cracking of medium- and high-severity is also larger shape, which indicates more variability in extending damage in the inspected sections. The overall value range is still limited, remaining below 10 m², which is tied to localized fading patterns for shorter road sections and since a recent inspection on the network.

In the case of total annual precipitation, mean annual temperature, and Average Annual Daily Truck Traffic (AADTT) the distributions are notably narrow and demonstrate low variability within the dataset, as evidenced in Figure 16. Total annual precipitation is centered around 1100 to about 1250 mm, implying even precipitation across the regions assessed. While mean annual temperature is similarly clustered about 13.2°C to roughly 13.8°C, indicating a similar

climate. On the other hand, AADTT presents slightly greater dispersion, with most values below 1500 trucks per day but a small portion approaching that upper bound.

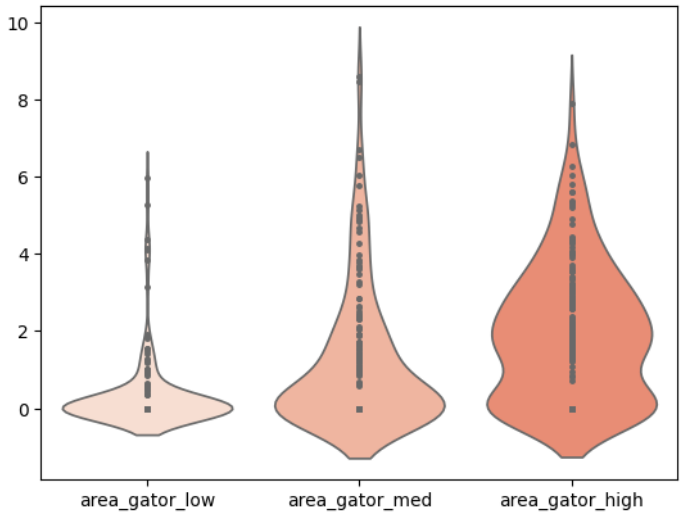



*Figure 15 Violin plot of alligator cracking areas for Italian Dataset*

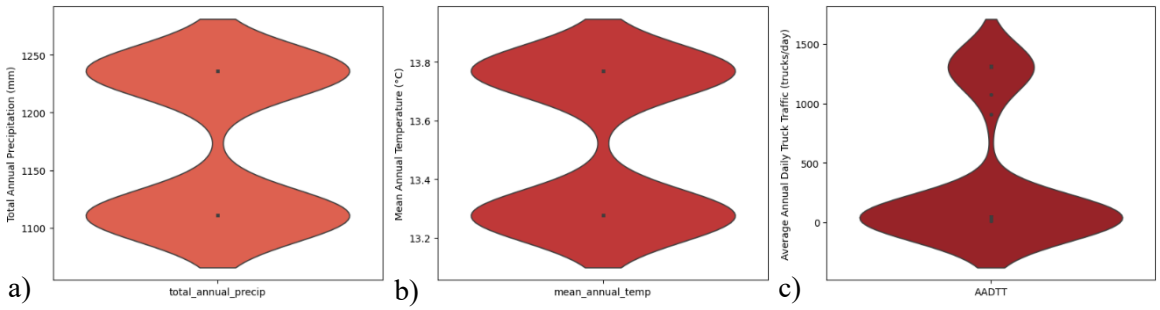



*Figure 16 Violin plots for Italian Dataset a) Total annual precipitation b) Mean Annual Temperature c) AADTT*

- **Bivariate analysis**

The correlation heatmap for the Italian dataset evidenced in Figure 17 displays a more heterogeneous correlation structure. The three severity levels of alligator cracking show moderate intercorrelations with low and high severity ($r = -0.24$) and with medium and high severity ($r = -0.53$), which may suggest a compensatory process in which an increase in one level of severity corresponds with a decrease in another, which is a possible outcome of the visual inspection classification process.

The AADTT variable shows a stronger positive correlation with low severity cracking ($r = 0.42$), suggesting that sections with heavier traffic loads have a tendency to develop initial damage at the surface. In the case of the environmental variables: total annual precipitation and mean annual temperature are almost perfectly and negatively correlated with an r =

−1.00, which indicates the limited variability and inverse seasonal relationship among both of these variables in the Italian sample.

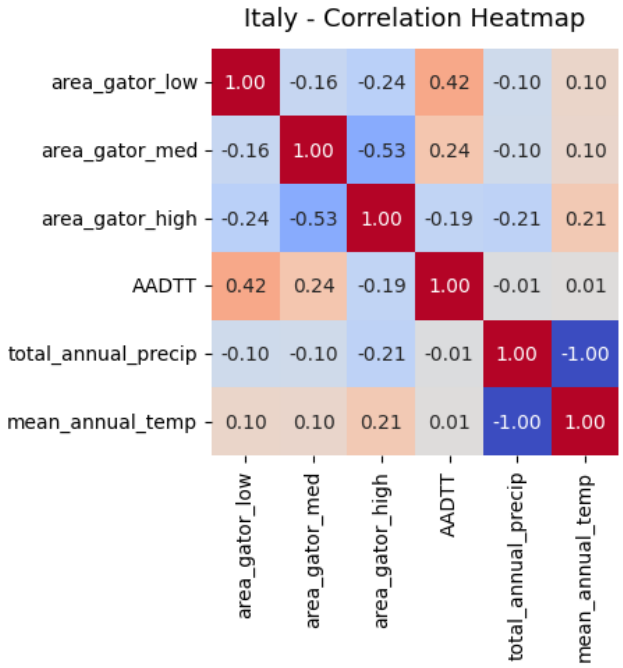



*Figure 17 Correlation heatmap of Italian dataset*

### 3.3.3 Comparison between datasets

- In contrast to the broad LTPP dataset, the Italian dataset is limited in scope and more uniform. It covers only two years (2024 – 2025) and one construction event, while the LTPP dataset consists of more than 11,000 observations across 49 U.S. states from 1980 – 2021, including multiple maintenance cycles. The mean areas of the deterioration levels in Italy are much lower (0.33 m² low, 1.18 m² medium, 1.92 m² high) than any of the corresponding deterioration levels in LTPP (17.34 m², 9.63 m², 7.24 m²), demonstrating limited degradation over a short monitoring time. Environmental conditions are also more homogeneous in Italy (≈13.5°C, 1173 mm) whereas LTPP covers a wide climatic range. Both datasets show significant variability in traffic characteristics, but the average AADTT in Italy is 362, which is significantly lower than AADTT in LTPP of 930. Overall, LTPP represents long-term, diverse deterioration patterns, while the Italian dataset offers a localized and recent snapshot of pavement conditions.

- When comparing the distribution of the alligator cracking areas between Italian and LTPP datasets, both reveal a right-skewed distribution pattern whereby the majority of sections fall within smaller damage areas, and decreasing percentage frequency with increasing damage area. However, the extent of deterioration varies significantly between the datasets. The LTPP dataset displays a wider range of damage area to several hundred square meters, reflecting the larger scale and greater range of heterogeneity in the U.S. pavement network. By comparison, the damage area for the Italian dataset, was limited to areas of less than 10 $m^2$ which can indicate that the sections surveyed in Italy were in better condition or in an earlier stage of deterioration.
- The violin plots of the datasets from the U.S. (LTPP) and Italy reveal marked differences in variability and distribution of environmental conditions as well as traffic conditions. For environmental conditions, the LTPP dataset depicts a broad range in total annual precipitation and in mean annual temperature, as this dataset represent different climates across several regions, some of which may be extreme. On the other hand, the Italian dataset generally has narrower distributions in both measures suggesting a more uniform environment with steady rainfall and moderate temperatures. In terms of traffic conditions, the LTPP dataset shows a very skewed AADTT distribution, where specific road segments exceed 8,000 trucks per day, showing that traffic intensity is highly heterogeneous across road segments. Alternatively, the Italian dataset shows a small AADTT variation, with most values being under 1,500 trucks per day, indicating that traffic is more uniformly lower in volume. In summary, this leads to the parallel conclusion that while the U.S. dataset captures more diverse ranges in climatic and operational conditions than the Italian dataset, the Italian dataset represents a more uniform setting, geographically and environmentally.
- After comparing the heatmaps of the two different datasets, it is observed that in the LTPP dataset the correlations between the variables tend to be weak, likely due to the considerable spatial and climatic diversity across the dataset, which minimizes the potential for observing linear relationships between all distress, traffic, or environmental factors. In contrast, the Italian dataset suggests stronger and clearer associations due to its small size and homogeneity. For instance, the nearly perfect negative correlation ($r = -1.00$) between mean annual temperature and total annual precipitation simply

exemplifies the local seasonal climate. Meanwhile, the moderate positive relationship between AADTT and low severity cracking suggests that heavier traffic is linked to earlier surface-related damage. Overall, these differences highlight how dataset scale and regional uniformity influence the observed statistical relationships.

## 3.4 Experimental Setup and Modeling Framework

### 3.4.1 Data Structuring and Preparation

#### 3.4.1.1 Definition of Subseries

To capture the temporal progression of deterioration, the data are organized into subseries, each representing the continuous evolution of pavement condition after a maintenance event. Both datasets contain the construction number variable which serves to distinguish maintenance or reconstruction activities on each road segment. Each time an intervention occurs, the construction number increases by one, and the alligator cracking areas referenced within each road segment resets back to zero, marking the beginning of a new pavement life cycle. The construction number thus allows tracking patterns of deterioration from the time a pavement section is renewed until the next maintenance action. Each subseries obtains its unique identity from the corresponding states, road segments, and construction number, represented in (1).

$$subseries = state\_name + segment\_id + cons\_number \quad (1)$$

Utilizing subseries is important for preserving the chronology of the data itself, and to ensure that deterioration models are written using data that has been captured under the same structural and operational conditions. This approach prevents the mixing of pre- and post-maintenance observations, allowing the model to learn deterioration dynamics that are both temporally coherent and physically meaningful.

#### 3.4.1.2 Data Division, Domain Standardization and Similarity Mapping

It is crucial in any predictive modeling task to split the data into training and testing subsets, as it allows for the estimation of the model's generalization ability—its capacity to perform accurately on new, independent data rather than memorizing patterns from the training set.

This is important because if the datasets are not divided, performance metrics could be artificially inflated, leading to overfitting and poor real-world applicability.

In this study, the data splitting procedure was designed to ensure methodological rigor and to avoid information leakage between the training and testing phases. The datasets were divided following a subseries-based approach rather than a random record-level split. As explained in section 3.4.1.1, each subseries represents a continuous temporal evolution of pavement condition for a specific road segment between two maintenance events, identified by the combination of state, segment, and construction number.

For the Italian dataset, 60 % of the subseries were randomly assigned to the training set thus 51 subseries and the remaining 40 % to the testing set (34 subseries), guaranteeing that all observations from the same segment remained within the same partition. This strategy prevents temporal or spatial overlap between training and testing samples, thereby ensuring that the model is evaluated on truly unseen data.

This split served as the foundation for the domain adaptation strategy. The 40% test set remained separate as the last measure of performance. The 60% training set measured a dual purpose: it was not only used in the final training dataset but was also the "template" for the K-Nearest Neighbors algorithm that allowed the model to find and learn from the most similar subseries within the larger US dataset in a way that augmented the training data with comparable, relevant examples from the source domain.

The division of 60/40 was chosen as a compromise between the learning capacity of the model and reliability of the evaluation. Given the limited temporal observations possible in each segment, this proportion of data provides sufficient subseries to allow for reasonably robust cross-validation while still providing a large enough and diverse testing set to develop a set of consistent and generalizable performance scores.

To validate the optimality of this choice, sensitivity analyses were conducted by over multiple train/test splits (90/10, 80/20, 70/30,60/40). For each split, the complete pipeline, including KNN selection and model tuning using Group K-Fold Cross-Validation was executed confirming that the final 60/40 proportion was judged to be the strongest, as it yielded the lowest Mean Absolute Error (MAE) on the independent test set.

Additionally, the training data was enriched with information from the United States Long-Term Pavement Performance (LTPP) database through a K-Nearest Neighbors

(KNN)–based selection process aimed at identifying U.S. subseries that exhibit environmental and traffic conditions similar to those in Italy. To achieve this, an embedding representation was built for both the Italian training subseries and all U.S. subseries, defined as the mean of total annual precipitation, mean annual temperature, and average annual daily truck traffic (AADTT) within each subseries as evidenced in equation (2).

$$e(sub) = (\overline{precip}\ (sub), \overline{temp}\ (sub), \overline{AADTT}\ (sub)) \quad (2)$$

Where:

- $\overline{\boldsymbol{precip}}\ (sub)$ is the mean total annual precipitation across all years of subseries sub.
- $\overline{\boldsymbol{temp}}\ (sub)$ is the mean annual temperature across all years of subseries sub.
- $\overline{AADTT}\ (sub)$ is the mean average annual daily truck traffic across all years of subseries sub.

Then, a standardization process was applied to ensure that the three variables used for the subseries embeddings were expressed on a comparable scale before computing Euclidean distances in the KNN similarity search. This procedure guarantees that the computed distances are meaningful and not biased by differences in variable magnitude or units. For this project, the StandardScaler [53] was chosen because it centers each variable by subtracting its mean and scales it to unit variance, thereby preserving the distribution shape while ensuring equal contribution of all features to the distance metric. In this implementation, the StandardScaler was fitted exclusively on the U.S. embeddings (source domain), meaning that the mean ($\mu_{US}$) and standard deviation ($\sigma_{US}$) were calculated using only the U.S. data.

Subsequently, both the U.S. and Italian embeddings were transformed using these same parameters and the equations (3)(4)(5). This approach ensures that the Italian data (target domain) is projected into the same standardized feature space defined by the U.S. domain, facilitating a fair and unbiased comparison of the two geographic regions. Importantly, by computing the normalization parameters solely from the source domain (U.S.), the procedure avoids any potential data leakage from the Italian dataset, maintaining the integrity of the transfer learning setup.

$$\widehat{precip}\ (sub) = \frac{\overline{precip}\ (sub) - \mu_{precip\ US}}{\sigma_{precip\ US}} \quad (3)$$

$$\widetilde{temp}(sub) = \frac{\overline{temp}\ (sub) - \mu_{temp\ US}}{\sigma_{temp\ US}} \quad (4)$$

$$\widetilde{AADTT}(sub) = \frac{\overline{AADTT}\ (sub) - \mu_{AADTT\ US}}{\sigma_{AADTT\ US}} \quad (5)$$

Where:

- $\overline{\boldsymbol{precip}}\ (sub), \overline{temp}\ (sub), \overline{AADTT}\ (sub)$ represent the components of the embedding vector.
- $\mu_{\boldsymbol{precip}\ US}$, $\mu_{temp\ US}$, $\mu_{AADTT\ US}$ denote the mean values of each corresponding variable calculated from all U.S. subseries, establishing the reference scale of the source domain.
- $\sigma_{\boldsymbol{precip}\ US}$, $\sigma_{temp\ US}$, $\sigma_{AADTT\ US}$ represent the standard deviations of those same variables across the U.S. dataset, used to normalize the spread of each feature.
- $\widetilde{\boldsymbol{precip}}\ (sub)$, $\widetilde{temp}(sub)$, $\widetilde{AADTT}(sub)$ are the standardized (z-score) values of each feature for subseries sub.

The standardized embeddings were then used to fit a KNN model on the U.S. subseries, which was then queried to find, for each Italian training subseries, its K most similar counterparts in the U.S. dataset using the Euclidean distance [53] evidenced in equation (6).

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (6)$$

Where:

- $n$ is the number of variables, in this case, 3: precipitation, temperature, and AADTT
- $x_i$ and $y_i$ are the standardized values of the $i$-th feature for the two points being compared.

The union of all identified neighboring U.S. subseries was collected to form the us_train pool. Finally, the selected U.S. subseries (us_train) were joined with the Italian training subset (it_train) to produce the "final training dataset" (train_all). This combined dataset gave

the model broader and more diverse representation of pavement deterioration patterns, enhancing its ability to generalize to the Italian context by utilizing structurally similar examples from the U.S. source domain. In contrast, the Italian test set remained completely separated and only used for evaluating performance to uphold the integrity of an independent testing phase.

### 3.4.2 Problem Formulation and Modeling Goals

#### 3.4.2.1 Predictive Task Definition

The predictive modeling task in this research is setup as a supervised multivariate one-step regression problem, focused on estimating the annual progression of pavement deterioration associated with alligator cracking on asphalt pavements across three severity levels: low, medium and high.

Given the limited temporal availability of Italian data, currently restricted to the years 2024 and 2025, the prediction is set to one year ahead. The multivariate nature of the task addresses the interdependence among severity levels, as low-, medium-, and high-severity cracking areas often evolve jointly through nonlinear deterioration processes.

For each subseries *i,* defined as a unique combination of state, road segment and construction number (see subsection 3.4.1.1), the objective is to predict the cracking area of severity level *s* at year *t,* denoted as $y_{i,t}^{(s)}$. This is achieved by learning a mapping function $f_s(\cdot)$ that relates both exogenous and endogenous explanatory variables to the target deterioration area.

- **Dependent Variables**

  The modeling framework aims to predict the annual extent of alligator cracking foreach severity level. Accordingly, three dependent variables were defined $y^{(low)}$, $y^{(med)}$, and $y^{(high)}$ each corresponding to the area of alligator cracking (in m²) of its severity level.

- **Independent Variables**

  The independent variables are categorized into two overall categories -exogenous and endogenous- that jointly capture the physical and temporal dynamics of pavement deterioration.

  - *Exogenous variables:* are external factors that affect the pavement but are not influenced by it such as the Average Annual Daily Truck Traffic (AADTT), Total

Annual Precipitation (Precip) and Mean Annual Temperature (Temp) which refer to traffic loading and climatic conditions observed at year *t.*

- *Endogenous variables:* represent the internal state of the pavement, captured through the lagged cracking areas of all severity levels from the previous year $A_{i,t-1}^{low}, A_{i,t-1}^{med}, A_{i,t-1}^{high}$.

- **Mathematical Formulation**

  The relationship between the dependent and independent variables can be expressed in the equation (7).

$$y_{i,t}^{(s)} = f_s\left(AADTT_{i,t}, Temp_{i,t}, Precip_{i,t}, A_{i,t-1}^{low}, A_{i,t-1}^{med}, A_{i,t-1}^{high}\right) + \varepsilon_{i,t}^{(s)}, \quad s \in \{low, med, high\} \tag{7}$$

  The function $f_s(\cdot)$ denotes the regression function learned for each severity level *s*, and $\varepsilon_{i,t}^{(s)}$ represents the model's random error term, capturing unobservable effects. The *i* identifies the subseries, while *t* denotes the year of observation.

- **Conceptual Integration**

  The formulation incorporates both exogenous influences (traffic and climate stressors) and endogenous deterioration mechanisms (prior cracking conditions), allowing the predictive model to account for nonlinear, dynamic, and interdependent processes controlling the development of pavement distress. The framework models these relationships with annual resolution and multiple severities and will enable data-driven forecasting of pavement cracking trajectories to support proactive maintenance planning and infrastructure management.

#### 3.4.2.2 Main Methodological Challenges

The model selection process was guided by several methodological and practical challenges identified during preliminary data analysis:

- **Limited Data in the Target Domain (t=2)**

  The Italian dataset contains only two temporal observations for every road segment, which severely limits the option of training data-hungry models. Limited data increases the risk of overfitting and limits the model's capability to learn temporal deterioration

behavior from only local data. Because of this, it becomes necessary to exploit the richer LTPP dataset as a complementary information source.

- **Domain Shift Between Source and Target Data**
  As evidenced in subchapter 3.3.2, the statistical properties of the independent variables (temperature, precipitation, traffic) as well as their relationship with the distress response differ significantly between the U.S. and Italian contexts. Such domain shift restricts the capability of a model trained solely on LTPP data, because it is unreasonable to apply the model to Italian conditions without modifications. Differences in climate zone and traffic create distributional shifts that need to be adjusted using domain adaptation or transfer learning techniques.
- **Multi-Severity Interdependence**
  The three categories of severity (low, medium, and high) are not independent, but instead represent a spectrum of deterioration. Capturing this interdependence might require multivariate modeling strategies that can jointly model the evolution across multiple severity classes.

### 3.4.3 Model Design and Selection

#### 3.4.3.1 Criteria for Model Selection

In order to respond to the issues identified above, —namely, limited data availability, domain mismatch between source and target datasets, and the inherent non-linearity of pavement deterioration processes—the predictive model selection was grounded in five main principles:

- **Performance on Tabular Data:** The modeling inputs—environmental variables, traffic variables, and historical deterioration measurements—are tabular data. Hence, models that have been demonstrated to work well on such data, namely tree-based ensemble models, are preferred based on their proven effectiveness and scalability.
- **Non- linearity and Interactions:** Pavement deterioration occurs due to complex, non-linear physical and environmental processes, where the interactions of various factors (e.g., traffic load, temperature, and precipitation) are crucial. Therefore, the chosen models are required to represent non-additive dependencies without requiring explicit manual specification of interaction terms.

- **Robustness:** Due to real-world infrastructure datasets being subject to significant measurement error, missing observations, and noise, robustness was another critical selection factor. Ensemble methods such as Random Forest, XGBoost, and LightGBM explicitly reduce variance through an averaging operation while improving stability to outliers and noisy observations.
- **Interpretability:** While predictive performance is essential, interpretability is still a major concern for engineering and decision-making purposes. Knowing the most influential factors in the deterioration process brings about actionable insights and corresponds to smart maintenance planning.
- **Suitability to Transfer Learning Situations:** Due to the short time span of the Italian dataset (2024–2025), the models are also tested for their potential to be used in cross-domain transfer learning scenarios. Instance-based methods like KNN exploit direct feature similarity between domains, while ensemble methods can be reweighted or fine-tuned to correct for distributional shifts, providing complementary strategies for knowledge extraction from richer source domains like the U.S. LTPP dataset.

In support of these principles, several modeling strategies were evaluated, including Random Forest, XGBoost, LightGBM and a K-Nearest Neighbors (KNN) instance-based transfer Learning approach. Additionally, a MultiOutputRegressor configuration was employed to model the interdependence between severity levels (low, medium, high) within a unified predictive framework. The intention of this comparative evaluation is to identify the model that appropriately balances predictive performance, interpretability, and transferability to another domain, supporting both methodological rigor and practical utility for data-deficient pavement management scenarios.

#### 3.4.3.2 Candidate Models

- **Classical Time-Series Models (ARIMA, PVAR)**

  These models were first considered due to the value of the PVAR to capture interdependencies between the three levels of severity. However, they were discarded because having t = 2 it is statistically impossible to estimate their parameters reliably.
- **Sequential Deep Learning Models (e.g., LSTM, GRU)**

Although powerful for sequences, the standard TL strategy (fine-tuning) is inapplicable. A re-training attempt on the target domain with a single sample per subseries would induce massive overfitting and forgetting of the source knowledge.

- **Baseline Model**

  In order to measure the improvements made by the forecasting models, a baseline model is necessary. For time-series forecasting problems, the most common basis is the naive model (or persistence model) [54], which assumes that the best prediction for the state at time *t* is simply the state observed at time *t-1*. This model does not include any explanatory variables as traffic or climatic variables, but relies completely on the concept of time-based persistence, which suggests that for deterioration, the most recent measurement is the best predictor of the next measurement.

  Formally, for a given subseries *i* and severity *s*, the naive prediction is expressed as:

$$\hat{y}_{i,t}^{(s)} = A_{i,t-1}^{(s)} \text{ for } s \in \{low, med, high\} \quad (8)$$

  Where $\hat{y}_{i,t}^{(s)}$ denotes the predicted deterioration area for subseries i at time t and severity level s, while $A_{i,t-1}^{(s)}$ is defined as the observed deterioration area from the previous year.

- **Random Forest (RF)**

  It is an ensemble ML model composed of combining several decision trees to achieve a more accurate and stable prediction. The term derives its name from producing a "forest" of randomly built decision trees [55]. The algorithm is a versatile design for many types of tasks, including both classification and regression.

  - **Performance on tabular data:** RF performs very well on structured tabular datasets with hardly any pre-processing and will naturally handle mixed scales of data types and feature distributional information.
  - **Non-linearity and interactions**: RF creates complex nonlinear relationships and high-order feature interactions automatically through its tree-based structure.
  - **Robustness:** Bagging (bootstrap aggregating) with random feature selection provides strong guarding against overfitting and outliers, regardless of noise or data availability, helping produce stable results when handling a noisy dataset or having limited data.

- **Interpretability:** RF can provide global interpretability through assessment of variable importance and local interpretability through partial dependence plots, promoting engineering understanding and insights into the mechanisms behind deterioration.
- **Suitability to transfer learning situations**: Due to the ensemble and stability of RF across domains or locations, it is a very versatile and adaptable machine learning model. It can generalize patterns learned in the source domain (U.S.) to the target domain (Italy) with minimal fine-tuning, especially when used with instance-based sampling from similar subseries.

- **XGBoost (Extreme Gradient Boosting)**

  It is an efficient and scalable version of gradient-boosted decision trees (GBDT). It constructs trees in succession, where each tree is constructed specifically to address the prediction errors of its predecessor(s)[56].

  - **Performance on tabular data:** It is usually labeled as state-of-the-art when it comes to predictive accuracy on tabular data.
  - **Non-linearity and interactions:** Strong capability to represent multi-dimensional, highly complex non-linearity and interactions through its methodical, gradient learning format.
  - **Robustness:** The sequentiality of boosting can introduce weaknesses regarding sensitivity to outliers and therefore robustness unless checked. However, it does come with L1 and L2 regularization parameters, which are powerful techniques used to inhibit overfitting and increase robustness.
  - **Interpretability:** Although more complex than RF, it offers some interpretability by measuring feature importance metrics, SHAP (SHapley Additive exPlanations) values, and gain for the variable contributions.
  - **Suitability to transfer learning situations**: The regularization and flexible structure of XGBoost provides utility in adaptation across domains, specifically when there is a change in the feature distribution between the source and target domains. Its strong generalization ensures reliable transfer when embedded with instance-based selection as KNN.

- **LightGBM (Light Gradient Boosting Machine)**

  It is a more recent framework based on GBDT, and is, on average, faster and more memory efficient than XGBoost. It grows trees lea1f-wise (best-first) rather than level-wise, which can result in faster convergence and improved accuracy [57].

  - **Performance on tabular data:** The performance is similar to or better than XGBoost, but with the added benefit of much faster training.
  - **Non-linearity and interactions:** It has the same excellent power as XGBoost for capturing complicated relationships. Leaf-wise growth may in some cases capture more complicated patterns.
  - **Robustness:** Like XGBoost, it includes regularization to help with overfitting. Leaf-wise growth is sometimes more likely to overfit with smaller datasets, but this is easy to manage with hyperparameter tuning (e.g., num_leaves).
  - **Interpretability:** Like XGBoost, LightGBM allows for ranking, or importance, and interpretability using SHAP, allowing for the ability to understand variable contributions across domains.
  - **Suitability to transfer learning situations**: Possesses the same high potential as XGBoost.

- **K-Nearest Neighbors (KNN) Instance-Based Transfer Learning**

  This is an instance-based learning algorithm that is non-parametric in nature. It creates access to domain adaptation through identifying, for each Italian subseries, the k most similar U.S. subseries based on standardized embeddings (precipitation, temperature, AADTT) [58].

  - **Performance on tabular data**: The KNN performance is dependent on a well-chosen distance metric and is sensitive to the “curse of dimensionality", where observations perform worse if there are more observations added that do not contain relevant information. Requires careful feature scaling.
  - **Non-linearity and interactions:** KNN performs well to learn complex and localized non-linear relationships under the assumption that similar inputs lead to similar outputs.

- **Robustness:** Despite its sensitivity to noise, KNN uses careful attention to scaling (StandardScaler) and optimal neighbor selection (K). KNN operates by averaging over neighboring observations, making local variability smoother, improving generalization.
- **Interpretability:** KNN is naturally interpretable - predictions can be traced backwards to observations or ensembles of specific source subseries back to origin.
- **Suitability for cross-domain transfer learning:** KNN was intentionally developed in this study as a transfer learning bridge by utilizing distance in the embedding space to find the closest U.S. instances relevant for making predictions about Italian instances. This logically combines the efforts and keeps domain correspondence without requiring retraining. Additionally, it serves as an interpretable data-based adaptation mechanism.

### 3.4.4 Performance Metrics

To assess the performance of the regression models, three standard error metrics were used: the coefficient of determination ($R^2$), the root mean square error (RMSE), and the mean absolute error (MAE). These metrics give various viewpoints on the predictive accuracy and reliability of the model.

- **Coefficient of Determination ($R^2$)**

  In this case, $R^2$ is utilized as an adjustment index comparing the model to the naive predictor that always returns the sample mean, rather than as a variance decomposition as is done with ordinary least squares and an intercept[59]. This interpretation is defined by equation (9), where *SSres* is the residual sum of squares and *SStot* is the total sum of squares. It holds for arbitrary predictive models and allows for below zero values when the model fits worse than the mean predictor. In the case of nonlinear and tree-based models using the classical "proportion of variance explained" interpretation has proved to be unreliable, so $R^2$ better regarded as a general goodness-of-fit measure, and is interpreted along with error metrics, such as MAE and RMSE.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{9}$$

- **Root Mean Square Error (RMSE)**

  The RMSE indicates the average difference between the predicted and actual alligator cracking areas, penalizing large errors more severely than smaller ones due to the quadratic term. A lower RMSE indicates higher model accuracy and better fit to the data [60].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

  It is computed using equation (10) where $y_i$ is the actual alligator cracking area of the *i-th* sample, $\hat{y}_i$ is the predicted alligator cracking area of the *i-th* sample and *n* is the number of samples.

- **Mean Absolute Error (MAE)**

  This metric denotes the average absolute difference between the actual alligator cracking areas and the alligator cracking area that was predicted. Unlike RMSE, the MAE treats all errors equally and does not penalize larger errors more. A lower MAE means that the model is more accurate [60].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (11)$$

  MAE is defined by equation (11) where $y_i$ is the actual alligator cracking area of the *i-th* sample, $\hat{y}_i$ is the predicted alligator cracking area of the *i-th* sample and *n* is the number of samples.

## 3.4.5 Hyperparameters selection

### 3.4.5.1 Randomized Search CV and Group K-Fold Cross-Validation

In order to obtain a valid measure of model performance, the cross-validation was based on the natural grouping of the data. Because observations within a subseries temporally and spatially correlated, it was important to ensure that no information leaks from the training subset to the validation subset of the cross-validation. For this reason, a Group K-Fold Cross-Validation strategy was applied where all samples from the same subseries remained together in a single fold. This assured the model was not trained and validated on a time period that

was overlapping for segments of the same road, allowing for an honest performance metric of the model generalization.

The cross-validation was implemented directly as part of the hyperparameter optimization process using the RandomizedSearchCV function in scikit-learn, it samples a fixed number of parameter combinations from the specified distributions, which has been shown to find very good combination of parameters in less time, especially when dealing with a large search space. This function was chosen over a traditional GridSearchCV because it is far more computationally efficient.

In this case, the tuning process involved a randomized search of 50 different combinations of hyperparameters for the models. In all separated evaluations for each hyperparameter combination, the GroupKFold procedure was performed with five folds. The number of folds was selected because it provides a robust and stable estimation of the model's true generalization error maintaining a computational efficiency. In each iteration, the model was trained on four of these folds (80% of the subseries) and validated on the one remaining fold (20 % of the subseries), which guaranteed validation on segments that were not seen during the training period.

The mean absolute error (MAE) served as a performance measure through the folds, and the average MAE determined the overall quality of each hyperparameter configuration. This metric was selected due to its robustness to outliers in comparison with the RMSE, which is affected more by large residuals.

The process of combining cross-validation with randomized hyperparameter search results in a strong and unbiased estimate of model performance, since the same model is trained and validated numerous times on different subsets of data. This effectively addresses overfitting to one subset and ensures performance remains consistent over folds, while the grouping method maintains temporal ordering and prevents leakage between correlated observations. Once the optimal hyperparameters—those resulting in the smallest average MAE—were selected, the model was trained again on the full training set to enable maximum learning while still maintaining the rigor and reliability achieved through cross-validation.

- **Random Forest Model (RF) with Transfer Learning (KNN)**

  For the Random Forest Model that includes transfer learning, the search space included parameters governing the shape of the tree, the sampling of features, and regularization at the nodes as summarized in Table 5.

*Table 5 Hyperparameter search space for Random Forest with transfer learning tuning.*

| Hyperparameter | Range explored | Description |
|---|---|---|
| **n_estimators** | [200, 400, 600, 800, 1000] | Number of trees in the ensemble |
| **max_depth** | [5, 9, 13, 17, 21, None] | Maximum depth of individual trees |
| **min_samples_split** | [2, 5, 10, 20] | Minimum samples to split a node |
| **min_samples_leaf** | [1, 2, 3, 5, 10] | Minimum samples at a leaf |
| **max_features** | [0.5, 0.75, 1.0] | Fraction or method for feature selection |

- **XGBoost Model (XGB) with Transfer Learning (KNN)**

  For the XGBoost Model that includes transfer learning, the search space included parameters to control model complexity, learning behavior, and regularization. The range explored for each hyperparameter is evidenced in Table 6.

*Table 6 Hyperparameter search space for XGBoost with transfer learning tuning.*

| Hyperparameter | Range explored | Description |
|---|---|---|
| **n_estimators** | [300, 600, 900, 1200] | Number of boosting trees in the ensemble |
| **max_depth** | [3, 4, 5, 6, 8] | Maximum depth of individual trees |
| **learning_rate** | [0.01, 0.03, 0.05, 0.1] | Shrinks the contribution of each tree |
| **subsample** | [0.6, 0.8, 1.0] | Fraction of training data used per tree |
| **colsample_bytree** | [0.6, 0.8, 1.0] | Fraction of features sampled per tree for diversity. |
| **min_child_ weight** | [1, 3, 5, 7] | Minimum sum of instance weights needed in a leaf node. |
| **reg_alpha** | [0.0, 0.001, 0.01, 0.1] | L1 regularization term |
| **reg_lambda** | [0.5, 1.0, 2.0] | L2 regularization term |

- **LightGBM Model (LGBM) with Transfer Learning (KNN)**

  The hyperparameter search for the LightGBM model with transfer learning concentrated on the aspects of controlling tree complexity, learning dynamics, and regularization force to ensure stable generalization across levels of severity. The ranges that were examined for each parameter, are provided in Table 7.

*Table 7 Hyperparameter search space for LightGBM with transfer learning tuning.*

| Hyperparameter | Range explored | Description |
|---|---|---|
| **n_estimators** | [400, 800, 1200, 1600, 2000] | Number of boosting trees in the ensemble |
| **max_depth** | [-1, 5, 7, 9, 11, 13] | Maximum depth of individual trees |
| **learning_rate** | [0.01, 0.03, 0.05, 0.07, 0.1] | Shrinks the contribution of each tree |
| **subsample** | [0.6, 0.7, 0.8, 0.9, 1.0] | Fraction of training data used per tree |
| **colsample_bytree** | [0.6, 0.7, 0.8, 0.9, 1.0] | Fraction of features sampled per tree for diversity. |
| **min_child_samples** | [5, 10, 20, 30, 50] | Minimum number of data points required in a leaf |
| **reg_alpha** | [0.0, 0.1, 0.5, 1.0] | L1 regularization term |
| **reg_lambda** | [0.0, 0.1, 0.5, 1.0] | L2 regularization term |
| **num_leaves** | [15, 31, 63, 95, 127] | Maximum number of leaves per tree |

#### 3.4.5.2 Number of Nearest Neighbors (K)

In this framework, the K-Nearest Neighbors (KNN) approach was utilized (see section 3.4.1.2) to find the most comparable pavement subseries from the U.S. dataset to each Italian subseries, creating the transfer set that was used in model training. The parameter K governs the number of neighbors that are included from the source domain. It can be thought of as a mediation between the correlation and diversity of transferred knowledge.

- If K is too small, the selected subseries can only include a very narrow piece of the source data, thus limiting variability and generalization capabilities of the model.
- If K is too large, the selected subseries can include dissimilar subseries that add noise and domain mismatch that reduces the prediction accuracy of the model for the end target domain.

An optimal K value (Kopt) was identified using a Group K-Fold cross-validation procedure on the training data generated from Italian training. For each candidate value of K, in a predefined grid (K=3, 5, 8, 10, 15, 20, 30), it was selected the closest U.S. subseries based on the Euclidean distance of features that had been standardized within the space of the features (precipitation, temperature, and AADTT). Each model was then trained and validated for each K, and performance was averaged across folds and the K that yielded the lowest MAE was chosen as Kopt ensuring a data-driven and robust choice that enhances the effectiveness of transfer learning.

### 3.4.6 Presentation of Graphical Results

Several plots are used to describe and analyze the results produced by the models. This subsection provides a detailed description of these plots by using some examples.

#### 3.4.6.1 Model Performance Overview

The predictive performance of each model is evaluated across the three severity levels (low, medium, high) and compared against the baseline model (see section 3.4.3.2), using the metrics of coefficient of determination ($R^2$), root mean square error (RMSE), and mean absolute error (MAE) metrics explained in the section 3.4.4.

The bar chart in Figure 18 presents an example of the coefficient of determination ($R^2$) values achieved by a model and the baseline model across three levels of pavement deterioration severity — Low, Medium, and High. The x-axis represents the severity levels, while the y-axis indicates the $R^2$ value. Violet bars correspond to the model on the training set, indicate how well the model fits known data; green bars represent the model on the test set and indicate how well the model generalizes to unseen data; and orange bars show the baseline mode, it is the benchmark for comparison; what can be obtained with a very simple approach.

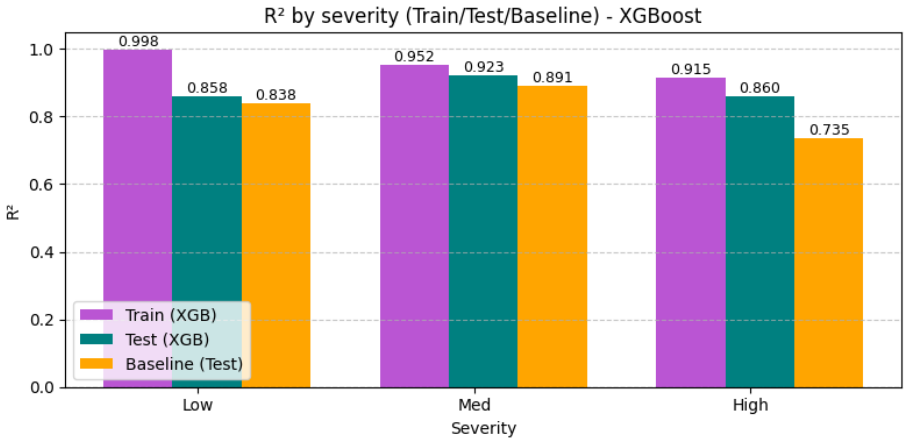



*Figure 18 Example of coefficient of determination ($R^2$) results by severity*

In Figure 19 and Figure 20 there are shown the bar charts of the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values for a model and baseline models in low, medium, and high levels of deterioration severity. The x-axis indicates the severity, which has three categories, Low, Med, and High. The y-axis shows the RMSE and MAE. The green

bars correspond to the model's RMSE on the test set and the orange bars correspond to the baseline model's RMSE on the test set**.**

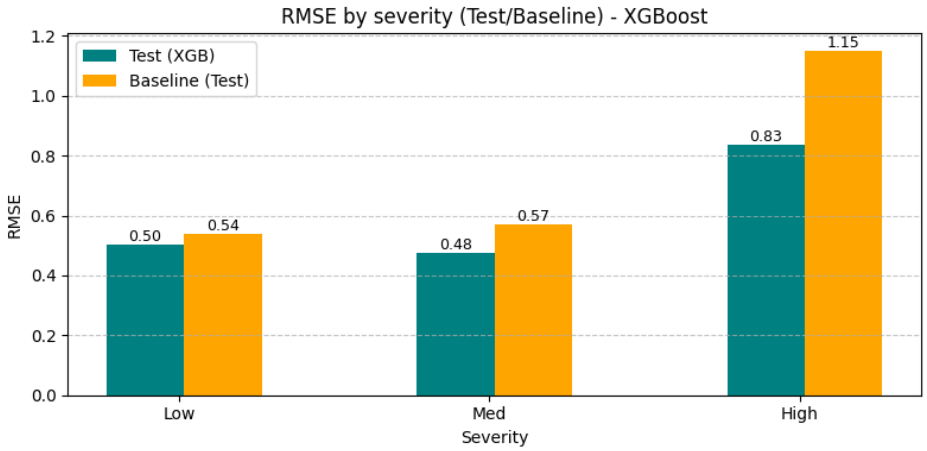



*Figure 19 Example of RMSE results by severity*

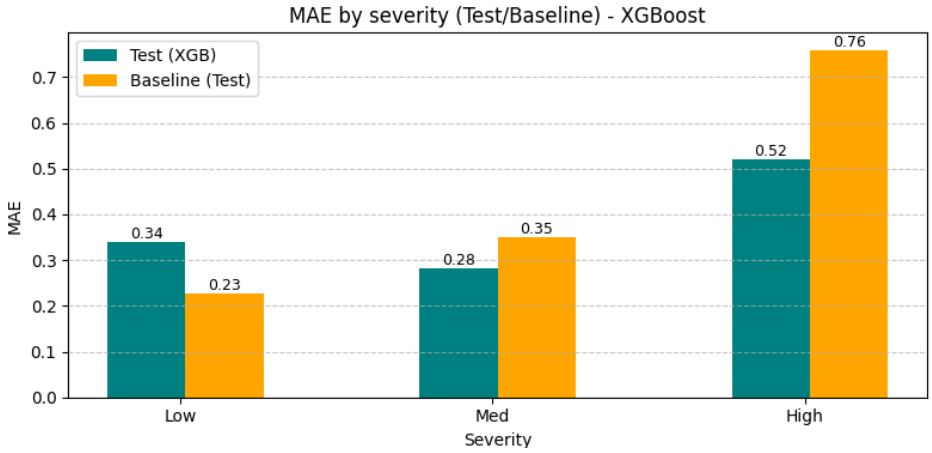



*Figure 20 Example of MAE results by severity*

### 3.4.6.2 Feature Importance Analysis

The bar chart in Figure 21 shows an example of the feature importance values of a model that was built to predict the variable area_gator_med (the area of alligator cracking in a medium severity state).

On the x-axis there is the importance score or numerical value of how much each input variable contributed to the predictions made by the model. On the y-axis are the input features selected in subchapter 3.4.2.1. The length of each horizontal bar represents the relative weight/influence of that feature in a prediction, highlighting which inputs are most informative for estimating the target output

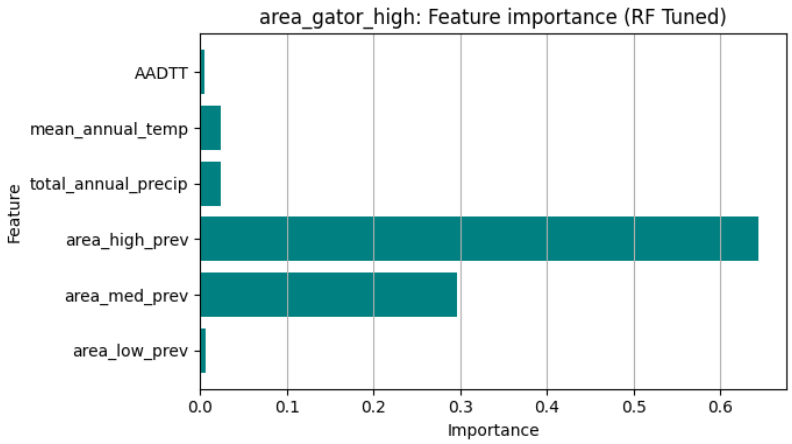



*Figure 21 Example of feature importance plot*

### 3.4.6.3 Predicted vs Observed comparison

The scatter plot in Figure 22 visualizes an example of the relationship between the observed and predicted high-severity alligator cracking area (area_gator_high) in the test data.

The x-axis depicts the actual area of deterioration, while the y-axis displays the values predicted through the model. The green circles represent predictions from the example model and the orange circles represent predictions from the baseline model. In addition, a black dashed line is provided to represent the perfect prediction reference in which predicted values are equal to observed values. The distance of the points from this line reflects the prediction error — the closer the points are to the line, the more accurate the prediction.

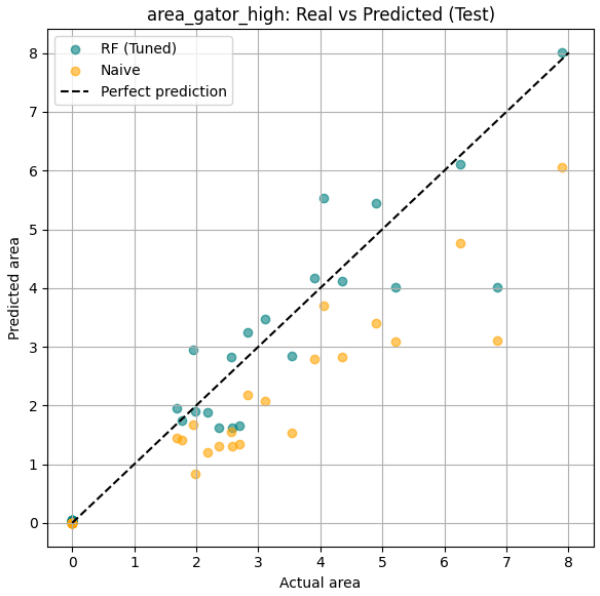



*Figure 22 Example of predicted vs observed plot*

### 3.4.6.4 Error Distribution Analysis

The Figure 23 illustrates the error distribution for a model and for the baseline model when predicting the medium-severity alligator cracking area (area_gator_med) on the test dataset. The x-axis shows the prediction error, calculated as the difference between the ground truth and predictive value (Error = y_real - y_predicted). This value reflects how far every prediction was from the actual value. Negative errors correspond to overestimations (predicted value was greater than the ground truth), while positive errors correspond to underestimations (predicted value was less than the ground truth). The y-axis depicts the frequency (number of test samples that fall within each error range).

Two overlapping histograms are shown: the green histogram represents the example model and the orange is the baseline model. The height of each bar indicates how many predictions generated an error value in that specific interval. This figure allows the reader to compare the shape, spread, and symmetry of the prediction errors from model to model, thus, evaluate which model's errors are the smallest and normally distributed around zero, suggesting better generalization to unseen data.

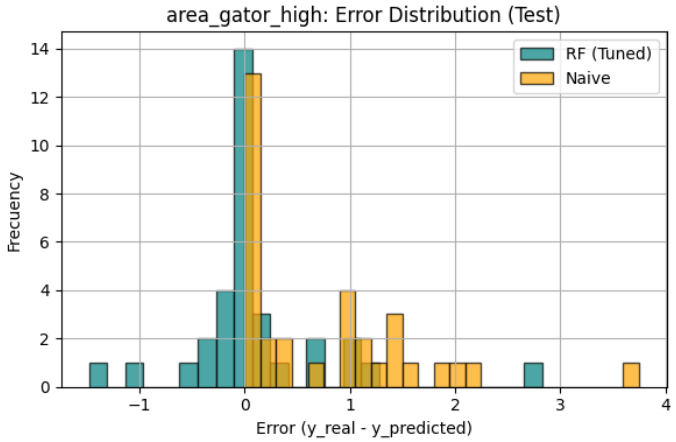



*Figure 23 Example of error distribution plot*

### 3.4.6.5 Learning Curve (MAE)

The line chart in Figure 24 presents the learning curve of a model for predicting the high-severity alligator cracking area (area_gator_high), using the Mean Absolute Error (MAE) as the evaluation metric.

The x-axis displays training size - that is, the number of training samples used in successive iterations to fit the model. The y-axis shows MAE (Mean Absolute Error) explained in section 3.4.4. Greater accuracy occurs when the MAE value is lower.

The violet line with circular markers indicates the CV training MAE, demonstrating how the model's error behaves on the data it was trained on as the sample size increases. The green line with square markers shows CV validation MAE, indicating the performance of the model on unseen data (validation folds) across different training sizes.

This type of graph is typically used to diagnosis learning behavior in models, helping to evaluate the performance of a model when data is increased in size, and to use inferences about its learning efficiency, bias–variance balance, and potential overfitting or underfitting behavior.

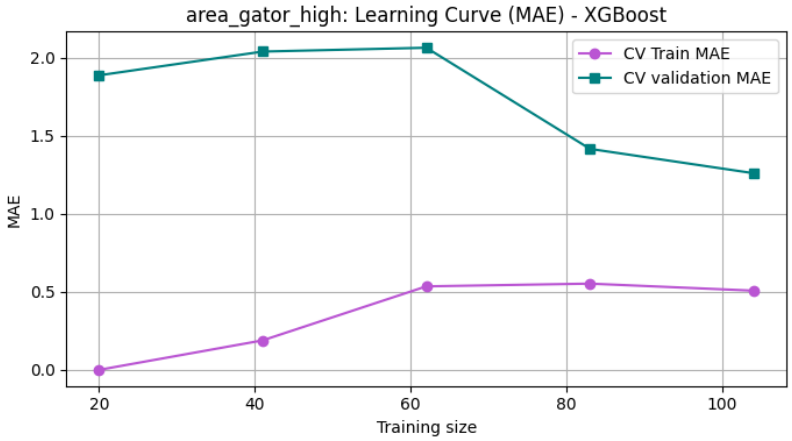



*Figure 24 Example of learning curve plot*

## 3.5 General Methodological limitations

### 3.5.1 Challenges in Severity Assessment

A labeling scheme was established for the purpose of classifying alligator cracking severities based on established civil engineering protocols. However, practical implementation revealed several constraints:

- Subjectivity: While the ASTM standards identify the criteria for assessment of severity, the practical ability to rate severity is consistently subjective and random depending on the professional.

- Redundant Frame Capture: One major limitation during the process was identifying the method of image capture. Since the imagery was collected from a camera facing forward to capture sequential images from the carriageway, it's possible that the same section of pavement was captured more than once. Therefore, labeling the severity is also redundant, whether similar areas are scoring variability based on strict uniformity or measuring independently through the sequential frames from such redundant observations.
- Perspective Distortion: All conversions for pixel to metric were based on a standard 3.5 meters lane width. However, variations in the camera angle and height may affect the geometric proportions of the pixel when relating it to the spatial measurements.
- A further complication stems from the inherent difficulty of reproducing identical imaging conditions during future data collection. With the original imaging not georeferenced or tightly controlled, the odds of replicating the exact spatial positions and camera orientations in future surveys is low.
- Pavement Structural Variability: Surveyed roads varied in design, not all of them are consistent with typical multilayer structurally asphalt pavement, particularly some rural roads, such as San Sebastiano da Po seem to have only the bituminous surface treatment (BST) layer and no structural base course. This variability may affect the form of surface cracking and obstruct the direct applicability of severity criteria developed for more robust urban pavements.

# 4

# RESULTS AND DICUSSION

This chapter provides a thorough analysis of the developed predictive models used in the deterioration problem of alligator cracking, investigating the individual and comparative performance of Random Forest, XGBoost, and LightGBM, each augmented with an instance-based transfer learning approach (KNN), at different severities of pavement deterioration (low, medium and high). Building on the methodological descriptions presented in Chapter 3, the outcomes are organized to emphasize not only the characteristics of each model, but more importantly, cross-model comparisons. The chapter begins by assessing each algorithm's predictive accuracy, generalizability, feature importance, and error structure for each severity level. It then synthesizes these findings to determine the best performing model within each level of pavement deterioration and discussing the implications for model stability, operational use, and pavement management.

## 4.1 Model Evaluation

### 4.1.1 Random Forest (RF) Applying Transfer Learning (KNN)

- **Random Forest Hyperparameters**

  The optimal hyperparameters shown in Table 8 and obtained through cross-validation explained in 3.4.5, were consistent across the three severity levels (low, medium, and high), suggesting that the Random Forest model converged toward a stable configuration regardless of the prediction target.

*Table 8 Optimal Random Forest hyperparameters after tuning.*

| Hyperparameter | Low severity | Medium severity | High severity |
|---|---|---|---|
| **n_estimators** | 800 | 800 | 800 |
| **max_depth** | None | None | None |
| **min_samples_split** | 2 | 2 | 2 |
| **min_samples_leaf** | 2 | 2 | 2 |
| **max_features** | 1 | 1 | 1 |

- **Selection of K in KNN**

  As explained in section 3.4.5, after applying the procedure and obtaining the results of Table 9, the optimal K value is 8. Although K = 5 yielded the lowest MAE (MAE = 0.302), the K= 8 achieved better $R^2$ value ($R^2$ = 0.806) and a comparable MAE (MAE = 0.304). Therefore, K = 8 was selected as the final configuration.

*Table 9 K values evaluated and its results*

| K value | MAE | $R^2$ |
|---|---|---|
| **3** | 0.321 | 0.685 |
| **5** | 0.302 | 0.765 |
| **8** | **0.304** | **0.806** |
| **10** | 0.321 | 0.743 |
| **15** | 0.319 | 0.751 |
| **20** | 0.325 | 0.727 |
| **30** | 0.546 | -1.141 |

- **Model Performance Overview**

The predictive performance results for Random Forest Model applying transfer learning are summarized in Figure 25, Figure 26 and Figure 27

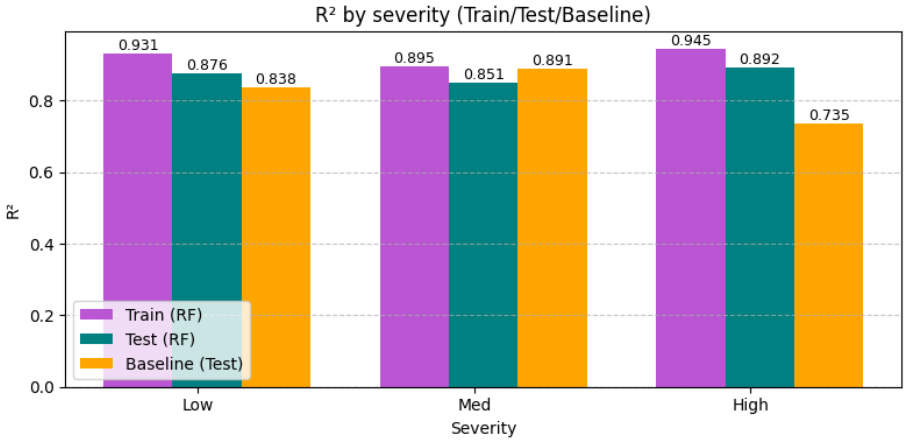



*Figure 25 Random Forest $R^2$ results by severity*

The $R^2$ values evidenced in Figure 25 show that the Random Forest offers a strong goodness-of-fit at all levels of severity. On the training set, the $R^2$ values range from 0.895 (medium severity) to 0.945 (high severity), indicating that the model does a good job of capturing the underlying relationships present in the data. The $R^2$ values on the test set are consistently high (0.876 for low, 0.851 for medium, 0.892 for high), indicating stable generalization. The narrow gap between train and test performance (within ±0.05 across all severities) suggests that the Random Forest captures the main nonlinear patterns without over-specializing to the training data without strong signs of overfitting.

When comparing to the baseline model, the Random Forest model had a consistently better performance for high severity events (0.892 vs. 0.735) with an implied increase of ≈0.16 of capturing the variability. While the baseline is still comparable for low severity (0.838) and medium severity (0.891), the Random Forest has a slight advantage, emphasizing the durability of the model across severity levels.

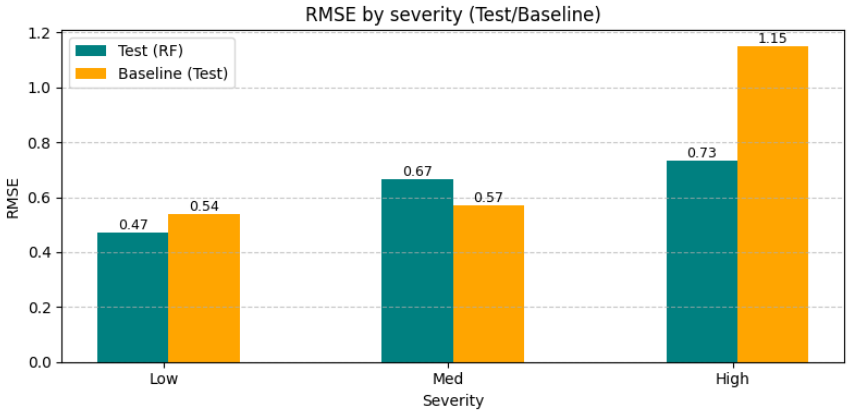



*Figure 26 Random Forest RMSE results by severity*

By looking at the error metrics (MAE and RMSE) for the RF model evidenced in the Figure 26 and Figure 27, in terms of severity levels, there is an obvious pattern. Both MAE (Low: 0.26, Med: 0.28, High: 0.42) and RMSE (Low: 0.47, Med: 0.67, High: 0.73) increase monotonically with severity. Essentially, tasks are more difficult to predict as severity increases. On average, the model's predictions remain less accurate and less precise with respect to variability for high-severity programs to those of low-severity reviews. The jump in RMSE from Low to Med (0.47 to 0.67) was especially pronounced, which is in agreement to the observed drop in performance observed in the $R^2$ metric per this group.

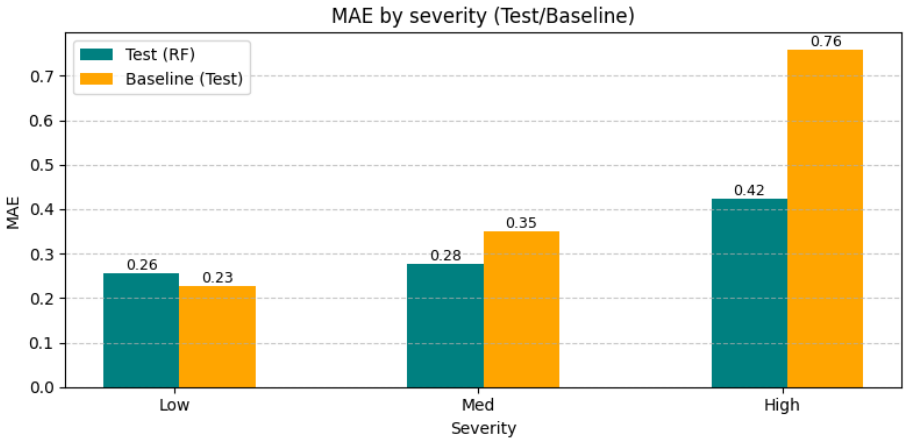



*Figure 27 Random Forest MAE results by severity*

Comparing with the baseline model, in the **low severity** group, the findings are somewhat mixed but directionally, the RF model is better. The RF model has a better fit ($R^2$: 0.876 vs. 0.838), and a lower RMSE (0.47 vs. 0.54) showing that it is less likely to make large errors. Conversely, the baseline model performs better on average (MAE: 0.23 vs. 0.26).

In the case of **medium severity** group, the baseline outperformed the RF on $R^2$ (0.891 vs. 0.851) and RMSE (0.57 vs. 0.67). The RF only slightly exceeded baseline on MAE (0.28 vs. 0.35). This performance indicates that while the RF's average error is lower it is a poorer-fit overall, while baseline is most likely to avoid large errors penalized heavily in RMSE metric for this subgroup.

Conversely, in the **high severity** model, the RF model has its greatest advantage where it clearly outperforms the baseline on every metric: a vastly greater $R^2$ (0.892 vs. 0.735), substantially less MAE (0.42 vs. 0.76), and considerably less RMSE (0.73 vs. 1.15). This means the RF model is much more capable than the baseline, and importantly (given the large

RMSE reduction), it is also better at avoiding large, erroneous predictions for this important subgroup.

- **Feature Importance Analysis**

The feature importance values derived from the tuned Random Forest models (see Figure 28, Figure 29, Figure 30) supply insights into the relative role of each predictor towards the estimation of deteriorated area by severity class (low, medium, and high). The analysis highlights a clear hierarchical structure among predictors, revealing that previous deterioration states (area_low_prev, area_med_prev, area_high_prev) play a dominant role in forecasting future conditions.

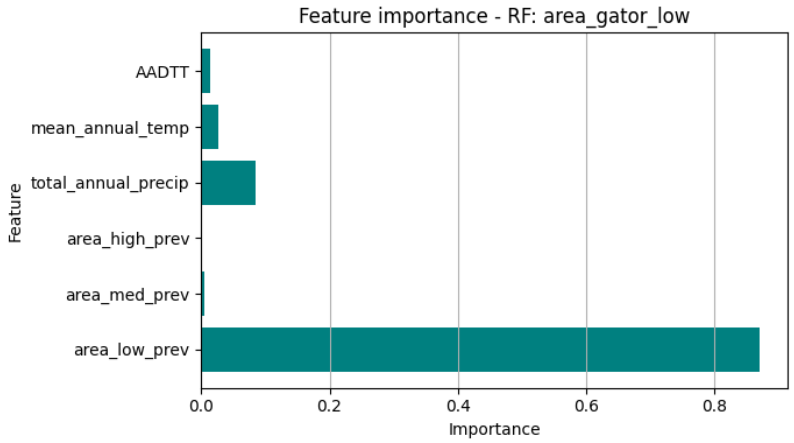



*Figure 28 Random Forest feature importance for low severity model*

For the **low-severity** deterioration model (Figure 28), the feature area_low_prev dominates the prediction almost completely, generating nearly all the explanatory power (importance ≈ 0.86), indicating that the extent of previously observed low-severity cracking is the most reliable predictor of its future progression.

Minor contributions of the climatic and traffic variables suggest that while having an environmental exposure may still modify the rate of deterioration, the short-term temporal persistence of low-severity cracking is primarily driven by its condition in the past.

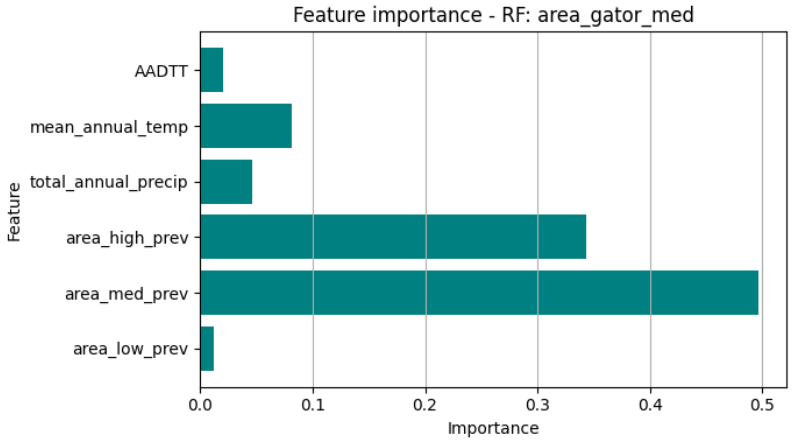



*Figure 29 Random Forest feature importance for medium severity model*

The **medium** severity model (Figure 29) exhibits the most complicated feature interaction. The previous medium severity area (area_med_prev) and previous high severity area (area_med_prev) are the most influential predictors (about 0.5 and 0.35), suggesting that the presence of medium severity deterioration tends to occur in previously medium and high severity cracking sections, indicating an evolution from mild to more severe pavement fatigue progression. The two potential points of dependence may be contributing to the model's unusual performance features (e.g., larger RMSE) in the previous model, as it is learning a more involved relationships involving a feedback loop that is nonlinear.

The increased relevance of mean annual temperature and total annual precipitation in this case, indicates the growing role of environmental factors as the pavement condition worsens and becomes more sensitive to external stresses.

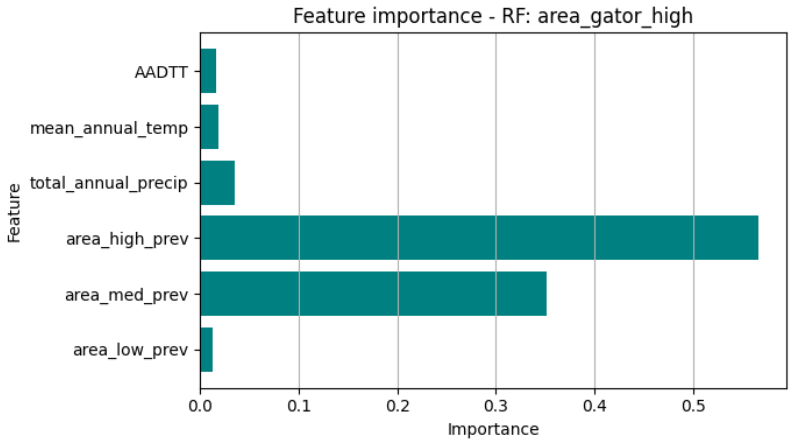



*Figure 30 Random Forest feature importance for high severity model*

The feature importance structure in the **high severity** model (Figure 30) shows a clear logic of persistence and escalation. There is an appreciable contribution from both area_high_prev (≈0.57) and area_med_prev (≈0.3), severe cracking evolves not only from its own prior extent but also from medium-level deterioration zones that escalate over time. Climatic and traffic variables (total_annual_precip, AADTT) remain low in relative importance while still increasing slightly above that of the lower severity models. This result suggest that once damage reaches a critical level, increased climatic and traffic loading promotes progression, whereas earlier time steps are predominantly shaped by prior conditions and persistence.

- **Predicted vs Observed comparison**

Figure 31, Figure 32 and Figure 33 illustrate the relationship between actual and predicted deteriorated areas for the Random Forest (RF) and the baseline (Naive) models.

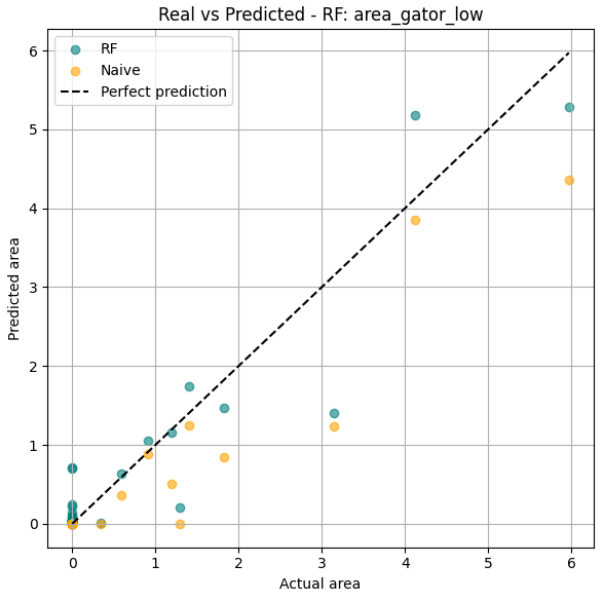



*Figure 31 Low severity predictive vs real values for Naive and Random Forest models*

In **low-severity** cracking, the scatterplot (see Figure 31) indicates a generally good relationship between the predicted and observed values, especially for smaller areas of deterioration. Most of the points are close to the origin, indicating that both models are able to characterize minor cracking patterns, which are the most typical in the data. However, the RF model displays a slightly closer clustering around the 45-degree line, indicating better predictability than the Naive approach, model that evidences under-prediction, particularly in actual areas in the range of 2 to 6.

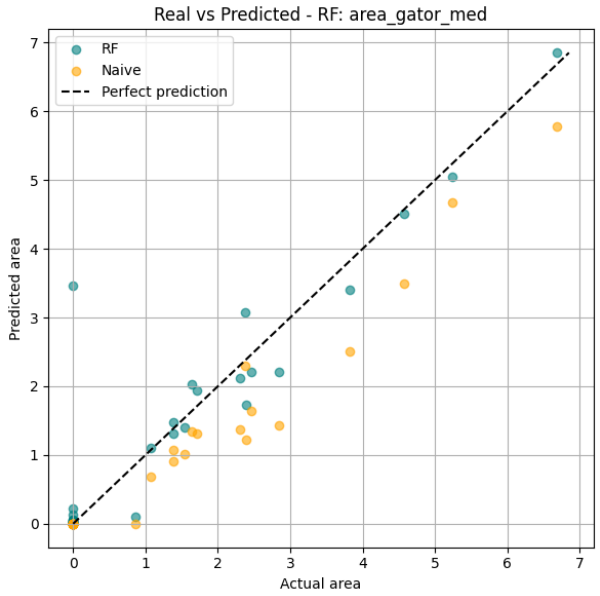



*Figure 32 Medium severity predictive vs real values for Naive and Random Forest models*

In the **medium severity** condition, the RF model reveals a strong relationship between predicted and actual areas (Figure 32), as the majority of the observations cluster around the 1:1 reference line. The slight vertical variability at higher actual values illustrates that some model uncertainty remains, however, the strong correlation trend exhibited validates the robustness of the model's performance and its ability to generalize the temporal progression of moderate severity cracks.

Even though the plot displays good predictive accuracy, the model performance overview evidenced poor behave in the $R^2$ and RMSE values. This may indicate that a few large prediction errors or limited variance in the observed data overly accounted for the numerical values. In this way, while the model depicts the overall trend well, its quantitative performance at this severity level remains watered down by outlier behavior and the non-uniform magnitudes of deterioration within this severity category.

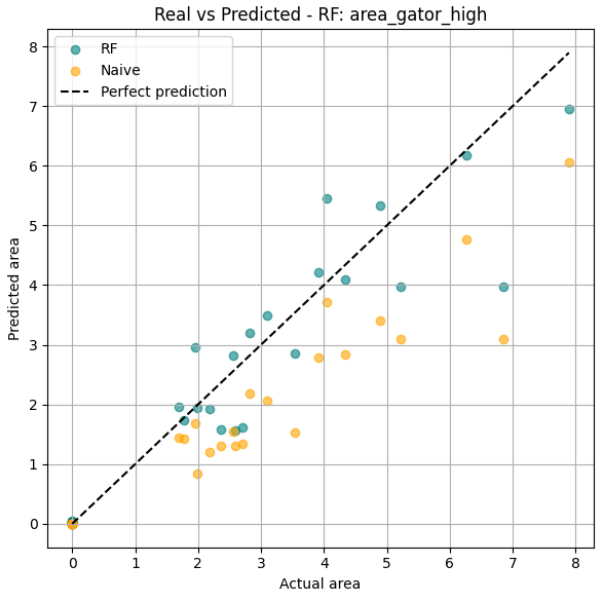



*Figure 33 High severity predictive vs real values for Naive and Random Forest models*

The **high severity** condition demonstrates the clearest benefit from the RF model over the baseline (Figure 33). There is a bit of dispersion for larger deterioration areas but the RF predictions are still consistently closer to the actual values than the Naive model predictions, which shows a clear tendency to underestimate the deterioration area. These performance results indicate that the RF model can adequately account for the cumulative and non-linear characteristic of more severe cracking processes, where large and divided damaged areas are likely to grow in a path dependent manner. The observed deviation at extreme values could be due to data sparseness since these types of high-severity areas are infrequent, but overall, the model generalizes their consistent behavior.

- **Error Distribution Analysis**

Figure 34, Figure 35 and Figure 36 present the distribution of prediction errors for both the tuned Random Forest (RF) model and the Naive baseline, across the three severity levels by showing the frequency with which each model makes errors of different magnitudes.

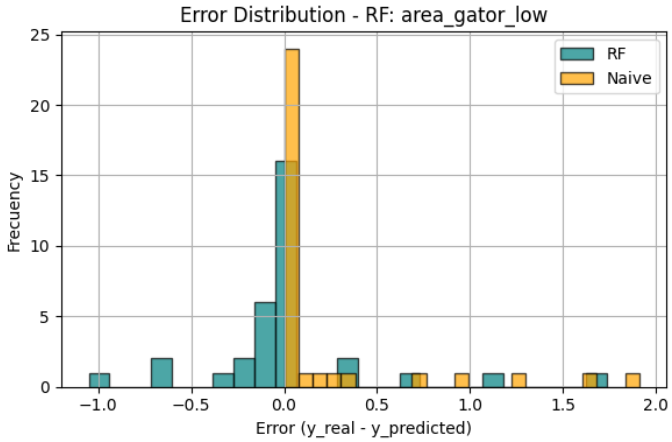



*Figure 34 Error distribution for the Random Forest model in low severity*

For the **low severity** case, the analysis of error distribution reveals contrasting patterns of behavior between the Naive and RF models, as evidenced in Figure 34. In the first one, errors are heavily concentrated at zero, indicating a high level of overall accuracy across many of the predictions, and a consistent tendency toward reproducing previous years' values with little deviation. However, the distribution contains notable asymmetry with a tail extending toward positive errors (1.5–2), indicating underestimation of actual deterioration in select instances.

In contrast, the RF model exhibits a wider and more dispersed error distribution compared to the Naive model. Although this model has a concentration of zero-error predictions, this frequency is notably lower. Its errors are more evenly distributed between --1 and 1.5 and some larger errors are also observed near 2.0. The RF modeling demonstrates more symmetry around zero and suggests that this model is less biased.

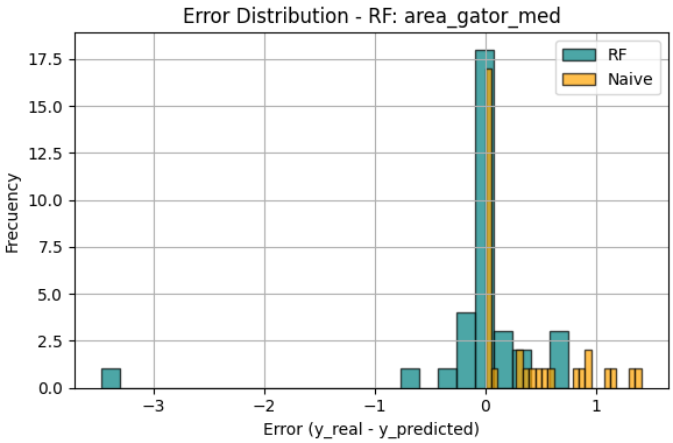



*Figure 35 Error distribution for the Random Forest model in medium severity*

When focusing on **medium severity** (see Figure 35), the RF model produces a narrow and centered error distribution predictions, with most values clustering closely around zero. This reaffirms that the model is generating accurate and consistent predictions for the majority of observations. By contrast, the Naive model generates a wider and more skewed error distribution with a long right tail, identifying it with an inherent bias in under-predicting deterioration. The single large negative error attributed to the RF model appear to be due to isolated cases of sudden deterioration not included in the training dataset, but unfortunately it penalizes the RF model in the RMSE calculation. Overall, the RF model exhibits less dispersion and a symmetrical presentation, validating its superior generalizability to the baseline method with less inference bias.

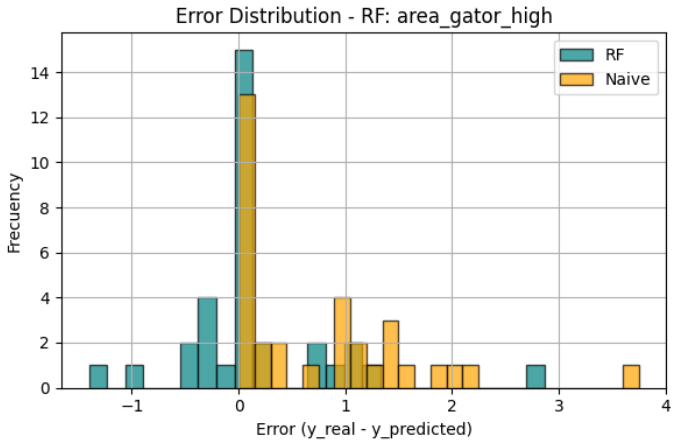



*Figure 36 Error distribution for the Random Forest model in high severity*

For the **high severity** model, the error distribution demonstrates an evident difference in predictive consistency between the RF methodology and the Naive approach. The Naive model shows a significantly wider and more asymmetric spread, reaching up to about +3.5. This long positive tail illustrates that the model has a systemic tendency to underestimate observed deterioration, meaning the Naive approach is often more likely to predict lower values than were actually observed. Such behavior suggests that the Naive approach struggles to capture the accelerated and nonlinear progression typical of advanced pavement failure, where damage propagation intensifies rapidly once the structural integrity of the material is compromised.

In contrast, the RF model has a more compact and symmetric error distribution that is centered around zero, with most residuals being from –1 to +1.2. This shape shows less bias and large errors than the Naive approach, and the small negative tail leads to the conclusion

that in some instances the RF model does slightly overestimate the observed deterioration, but on aggregate, the residuals are centered around zero.

- **Learning Curve (MAE)**

Figure 37, Figure 38 and Figure 39 present the learning curve for the low, medium and high severity models.

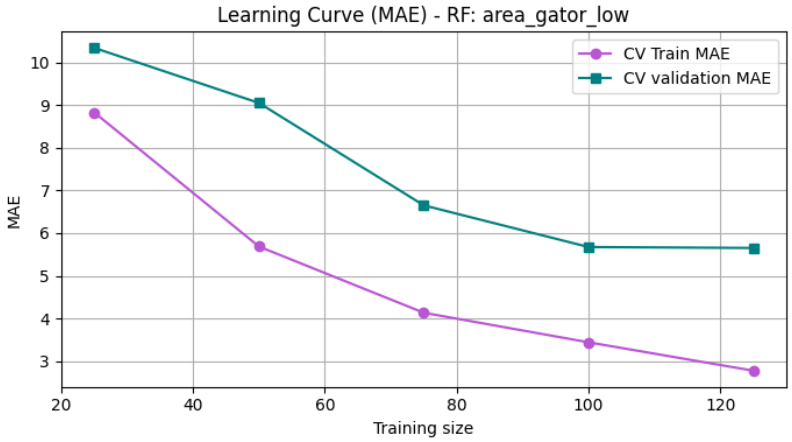



*Figure 37 Random Forest Learning curve for low severity model*

The learning curve for the **low severity** model (Figure 37) shows a high-variability, or overfitting, pattern. Both, the validation and training errors steadily decreased with increasing sample size. In the case of training curve, it reached a low MAE of around 2.8, which demonstrates that the model learned the training data very well. Comparatively, the validation error remained much higher overall sample sizes, and ultimately plateaued near 5.6 MAE. The large and sustained gap between the two curves demonstrates that the model did not generalize well to unseen data, likely indicating it has not learned the underlying patterns of deterioration, and instead, partially memorized the noise and details of the training dataset. The flattening trend in validation curve also suggests that simply adding more data won't be sufficient to close the gap. Rather, in order to improve the generalization, it is necessary to consider reducing model complexity with additional regularization or enhanced feature representation to help reduce overfitting and achieve a stronger generalization performance.

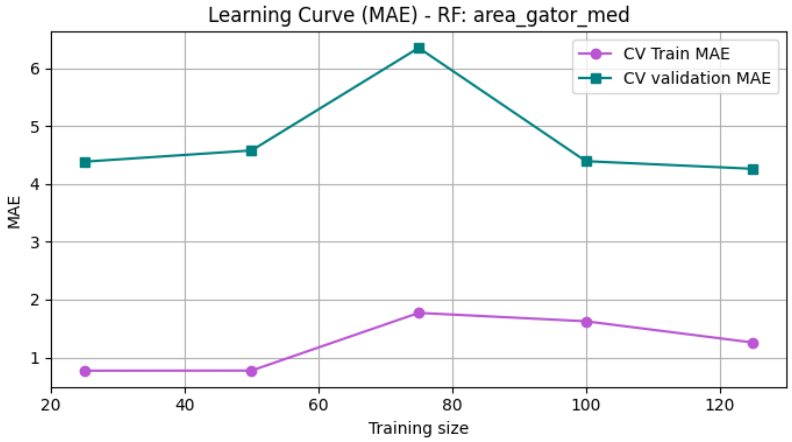



*Figure 38 Random Forest learning curve for medium severity model*

For the case of **medium severity**, the learning curve (Figure 38) it shows a relatively unstable pattern. The training MAE is low for all training sizes but the validation MAE shows marked variability, including a spike for samples at the mid-range training sample size. This volatility in performance signifies differences in model performance depending upon the specific data used for validation which indicates high variability in data, or outliers that determine validation error. Additionally, there is a noticeable gap between the two curves which suggests overfitting behavior wherein the model is trained on the training dataset and are not able to generalize. This behavior may stem from the limited representation of medium-severity cases or from the heterogeneous nature of this category, where transitions between low and high severity create mixed signal patterns. Increasing the sample size, introducing regularization, or incorporating more representative features could help smooth these fluctuations and enhance generalization stability.

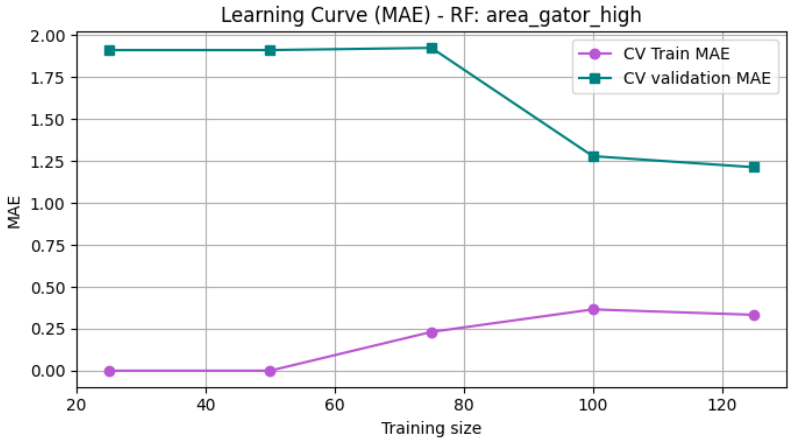



*Figure 39 Random Forest learning curve for high severity model*

The learning curve of **high severity** model shows that at smaller training sizes, the training MAE is extremely low (nearly zero), while the cross-validated MAE is much higher (around 1.9) indicating preliminary overfitting — the model learns the limited training examples almost perfectly, but performs poorly on unseen data. However, as the size of the training increases, both curves move closer together: the validation error decreases, and the training error makes a small increase up to around 0.3. This is typical and healthy behavior for a well-learning model.

The consistent downward trend of the validation MAE indicates that the Random Forest model benefits from having larger training sets and that its predictive performance for high-severity cracking continues to improve with additional data. Additionally, the convergence of both curves toward lower MAE values also means that variance has been reduced and the model has improved stability, implying that the model has captured the main patterns in the data without overfitting excessively.

In summary, this curve reflects effective learning and good potential for generalization for the high severity, where increasing the training data size leads to progressively better and more reliable predictions.

- **Discussion and Interpretation**

This thorough evaluation indicates the speculative performance of the Random Forest (RF) model is highly variable and dependent upon the specific severity class. The evaluation suggests that the model is not simply better or worse; it instead ranges from distinctly

excellent to inaccurately poor depending on the complexity and nature of the prediction task associated with each severity level.

For the high severity class, the model performs clearly best. It consistently outperforms the baseline along all metrics, and most clearly on $R^2$ (0.892 vs. 0.735) and RMSE (0.73 vs. 1.15). The feature importance analysis clearly supports this, revealing that the model considers previous levels of persistence (area_high_prev) and even escalation (area_med_prev). The error distributions show that the RF model is unbiased, with errors centered at zero, while the Naive baseline shows a systematic under-prediction bias. Furthermore, the learning curve for this model is also healthy, showing convergence, suggesting that predict rate would be improved with increased training data all of which supports the claim.

Conversely, the model specific to medium-severity deterioration exhibits a notable, somewhat surprising anomaly, as it performs worse than the baseline on key performance metrics ($R^2 = 0.851$ compared to 0.891 and RMSE = 0.67 relative to 0.57). This discrepancy is unusual, as it may be a reflection of the model encountering a particularly difficult interaction of features where it must learn simultaneously from two different paths of deterioration (i.e., persistence, area_med_prev, and de-escalation, area_high_prev). This overlapping of dependencies creates a structural instability, as evinced by the error distribution, which contains one large negative residual that punishes the model's RMSE through squared error accumulation. Furthermore, the learning curve reflects this instability as there are sharp variances and spikes in the learning curves instead of convergence, which suggests a heightened sensitivity to the respective data subsets used for training and validation.

Likewise, the low-severity model seems to be overfitting, probably based on complexity of model not matching simplicity of task. The feature importance plot shows that prediction in this regime can almost entirely be attributed to the area_low_prev variable, as it carries a value of about 86% of the total implicitly weight. This indicates that the modeling process is being conducted by a very simple persistence process, which needs not the representational depth of an ensemble model such as the Random Forest. The corresponding learning curve supports this interpretation, showing typical high-variance signs of overfitting where the

model is able to obtain a very low training error but does not generalize to new data effectively.

In conclusion, the efficacy of the Random Forest (RF) model using transfer learning (KNN) seems to be highly context-dependent. It is substantially more valuable when modeling the complicated, nonlinear properties of the high-severity regime, in which its ensemble model captures the more complicated structures of deterioration. However, it does not perform well when modeling the more unstable and dual-path interactions associated with the medium-severity condition, summing the fact that performance could have been penalized by the presence of few influential outliers that affected the error-based metrics like RMSE. In the low-severity regime, the RF model exhibits tendencies toward overfitting, as its own complexity is greater than the simple persistence-driven relation it needed to model.

### 4.1.2 XGBoost (XGB) Applying Transfer Learning (KNN)

- **XGBoost Hyperparameters**

  Following Randomized Search with Group K-Fold cross-validation explained in section 3.4.5, the best-performing configuration of hyperparameters was found for each set of severity levels. The results, presented in Table 10, revealed that a unique set of hyperparameter values is required for each of the three target classes: low, medium, and high severity. This suggests that model structure must be specialized to effectively capture the different relationships in the data across each target variable.

*Table 10 Optimal XGBoost hyperparameters after tuning.*

| Hyperparameter | Low severity | Medium severity | High severity |
|---|---|---|---|
| **n_estimators** | 300 | 300 | 300 |
| **max_depth** | 4 | 8 | 4 |
| **learning_rate** | 0.03 | 0.01 | 0.01 |
| **subsample** | 1 | 1 | 1 |
| **colsample_bytree** | 1 | 1 | 1 |
| **min_child_weight** | 1 | 5 | 5 |
| **reg_alpha** | 0.1 | 0.01 | 0.1 |
| **reg_lambda** | 2 | 2 | 2 |

- **Selection of K in KNN**

  As explained in section 3.4.5, various K values were analyzed to identify the most appropriate number of neighbors for the KNN-based instance selection. The results indicate that although K = 3 resulted in the lowest MAE (MAE = 0.345), K = 8 performed almost equally as well (MAE=0.354) with a higher $R^2$ ($R^2$ = 0.298) showing that K = 8 would yield slightly more stable and accurate predictions with adequate model generalization. Thus, K=8 was selected as the ideal value for future analyses.

*Table 11 K values evaluated in XGBoost and its results*

| K value | MAE | R2 |
|---|---|---|
| **3** | 0.345 | 0.239 |
| **5** | 0.364 | 0.342 |
| **8** | **0.354** | **0.298** |
| **10** | 0.370 | 0.282 |
| **15** | 0.389 | 0.380 |
| **20** | 0.428 | -2.765 |
| **30** | 0.519 | 0.229 |

- **Model Performance Overview**

The predictive performance results for XGBoost (XGB) model across the three severity levels are shown in Figure 40, Figure 41 and Figure 42.

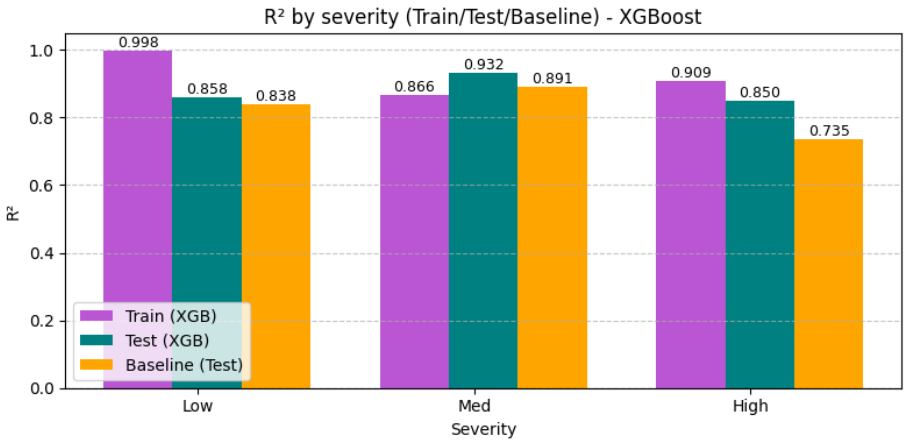



*Figure 40 XGBoost $R^2$ results by severity*

The integration of XGBoost and KNN-based instance selection delivered consistently good predictive performance across all severity levels (see Figure 40). Regarding the goodness-of-fit, the XGBoost model shows substantial predictive power across all severities, with $R^2$

values ranging from 0.866 up to 0.998 for train and values from 0.850 up to 0.932 for test case. For low severity the model achieves near-perfect training fit ($R^2$ = 0.998) but a much more moderate test performance ($R^2$ = 0.858), only very slightly better than the baseline ($R^2$ = 0.838). This indicates overfitting in the lower range, where variability in deterioration is naturally low. At medium severity, the model achieves lower generalization ($R^2$ train = 0.866; $R^2$ test = 0.932) outperforming the baseline ($R^2$ = 0.891) and confirming its ability to model non-linear dependencies between climate, traffic and historical deterioration. For high severity, XGBoost continues to outperform the baseline ($R^2$ test = 0.850 vs. 0.735), highlighting its strength in capturing threshold effects and accelerated deterioration typical of advanced cracking stages.

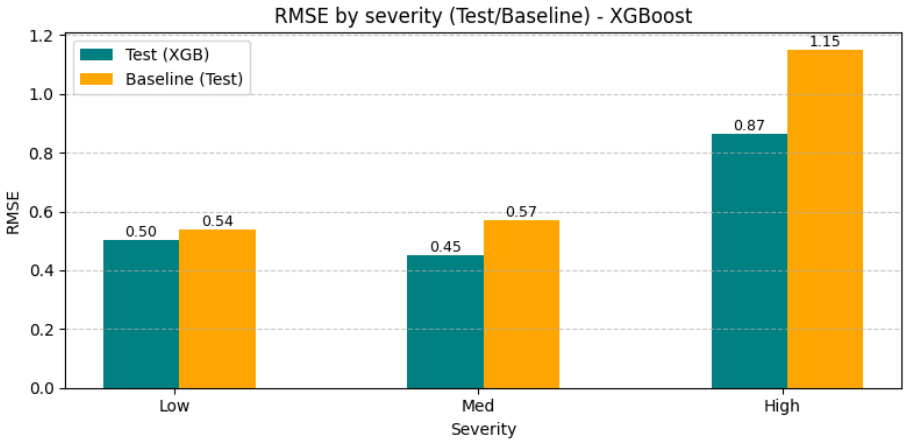



*Figure 41 XGBoost RMSE results by severity*

According to the Root Mean Squared Error (RMSE) in Figure 41, the XGBoost model demonstrates better predictive accuracy across all levels of severity. It consistently achieves better RMSE values (Low: 0.50, Medium: 0.45, High: 0.87) than the baseline model (low: 0.54, medium: 0.57, high: 1.15). The greatest improvement takes place in the high severity level where XGBoost reduces its RMSE by roughly 24% from baseline. This indicates significantly better capacity to handle complex prediction tasks and to restrain large deviations from the predicted value, which is deemed heavily through the RMSE.

Conversely, the Mean Absolute Error (MAE) in Figure 42 shows a more complex story. Different from RMSE, the baseline model achieves lower (and thus better) MAE values at the low (0.23 vs. 0.34) and medium (0.35 vs. 0.38) severity level while XGBoost is only better at predicting the high severity case (0.54 vs. 0.76). This disjunction of RMSE and MAE results suggests that while the XGBoost model is better at diminishing large prediction

errors (congruent with its best RMSE), its average deviation at predicting low and medium severity was higher, on average, than the baseline. That said, the high severity level remains the area in which the XGBoost promotes a discernible advantage, resulting in strong predictions in both RMSE and MAE.

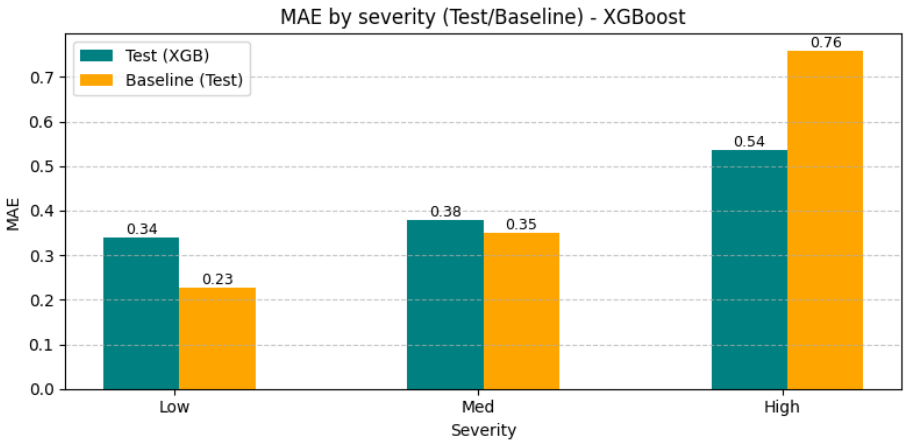



*Figure 42 XGBoost MAE results by severity*

Comparing with the baseline model, with regard to **low-severity** scenarios, results are somewhat mixed. The baseline model is better on average, it yields a lower MAE of 0.23 compared to the 0.34 obtained by XGBoost. However, the XGBoost model does slightly better at avoiding large errors—that is, with a lower RMSE of 0.50 versus 0.54 but unfortunately it shows the most significant signs of overfitting; that is, 0.998 $R^2$ on train versus 0.858 $R^2$ on test.

In the **medium-severity** group, the baseline model again has a slight advantage for average error (0.35 versus 0.38). However, the XGBoost model is clearly superior at minimizing large errors, with RMSE of 0.45 (versus the baseline at RMSE of 0.57), and it does a better job capturing the variability in the data (0.932 versus the baseline at 0.891).

The **high-severity** category is the clear strength of the XGBoost model. It outperforms the baseline in all three metrics: average error (MAE of 0.54 versus 0.76), large errors (RMSE of 0.87 versus 1.15), and providing a much better model fit ($R^2$ of 0.850 versus 0.735). This means that the model is probably most valuable when predicting the most critical outcomes.

- **Feature Importance Analysis**

An analysis of feature importance was conducted using the tuned XGBoost models (Figure 43, Figure 44 and Figure 45) to quantify the relative influence of each predictor on the estimated deteriorated area per severity class.

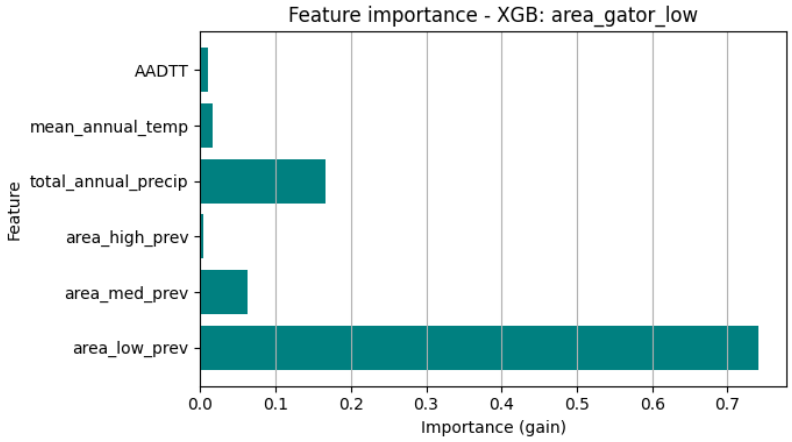



*Figure 43 XGBoost feature importance for low severity model*

For **low-severity** deterioration, the area of low-severity cracking in the prior year (area_low_prev), dominates the model by far contributing to over 74% of the total predicted gain, indicating that the existing condition itself is the greatest prerequisite for later deterioration. However, the second most dominant attribute is a key environmental factor: total_annual_precip (contributing almost 16%) specifies that while area_low_prev is a clear indicator of the potential for deterioration, precipitation is the main external forcing element driving deterioration, which is consistent with engineering principles. At this moment, it is early enough, the pavement is still structurally sound, then when the water infiltrates the cracks, it freezes and expands acting as a wedge expanding the existing cracks and enabling crack propagation that causes the area deterioration.

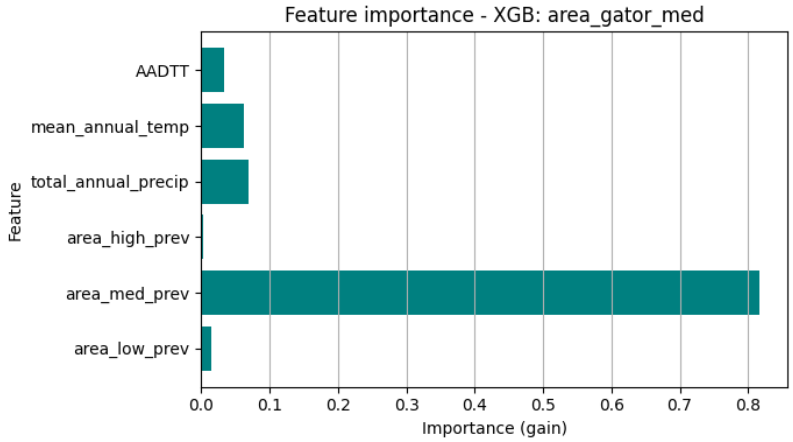



*Figure 44 XGBoost feature importance for medium severity model*

In the **medium-severity** model (Figure 44), the variable area_med_prev is by far the strongest predictor, with an importance score of about 0.82 indicating that the extent of previously observed medium-severity cracking is the most reliable predictor of its future progression. In contrast, the climate and traffic variables contribute only marginally to the prediction.

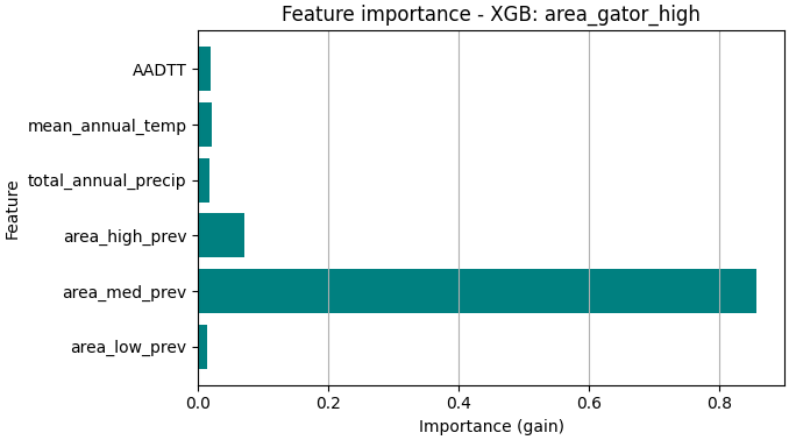



*Figure 45 XGBoost feature importance for high severity model*

In the case of the **high-severity** deterioration, the feature importance graph presented in Figure 45, reveals the area_med_prev as the overwhelmingly dominant predictor, contributing approximately 86% of total gain. On the other hand, the variables related to climate and traffic, have weights that are close to zero. Although this seems inconsistent with engineering principles that establish them as the key physical accelerators of deterioration, it likely reflects how the model interprets the data rather than an absence of causal influence. Two main explanations support this interpretation:

- Structural irreversibility: The model may consider area_med_prev a structural "point of no return" that once the cracking reaches this point, the pavement integrity is already compromised and it is certain to go on to high severity, making area_med_prev a much stronger predictor than the original, causal predictors.
- Proxy limitations: Input variables like total_annual_precip are low-quality proxies for the actual failure mechanisms and it does not account for specific events that are damaging (e.g., severe storms, freeze-thaw cycles), and therefore the model assigned zero predictive weight.

Taken together, the analysis of feature importance for all three severity indices delivers a compelling narrative for the dynamic life cycle of pavement failure. The low severity model demonstrates an "environmental initiation" phase in which damage progression is driven by the existing state (area_low_prev), as well as an external factor, total_annual_precip, that likely induces the freeze-thaw cycle. Then the mechanism transitions in the medium severity model to a more complex "structural failure" phase, which identifies the presence of area_high_prev to be the most predominant predictor of medium severity, suggesting a compromised pavement base is the primary driver for medium severity. Finally, the high severity model displays the "inevitable collapse" phase; a simple, linear progression of damage and singular predictor, area_med_prev, alongside original drivers such as climate and traffic that become irrelevant. The strong transition from an environmental driver, to a structural driver, then to an inevitable mechanism of failure, strongly justifies the need for three distinct models.

- **Predicted vs Observed comparison**

The relationship between real vs predicted deteriorated areas for the XGBoost (XGB) and Naive models, for each severity class is shown in the Figure 46, Figure 47 and Figure 48.

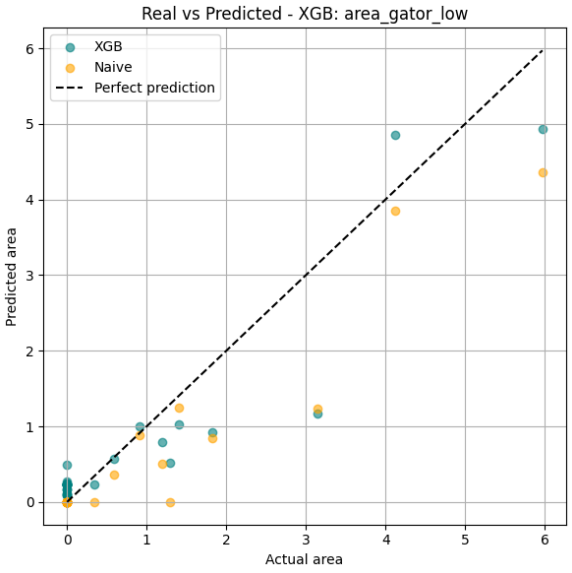



*Figure 46 Low severity predictive vs real values for Naive and XGBoost models*

For **low-severity** deterioration areas, as evidenced in Figure 46, most of predicted values from the XGBoost model reside very close to the 1:1 line, indicating a good fit between predicted versus observed areas. The XGBoost model predictions also exhibited slightly lower errors and bias in comparison to the Naive baseline predictions, especially for the lower predicted areas of deterioration, whereas Naive baseline approaches tended to underpredict observed low severity deterioration areas. In both cases, the models tended to diverge a bit more at the higher observed areas (>2 m²), since there were fewer observations for larger magnitude locations within the dataset.

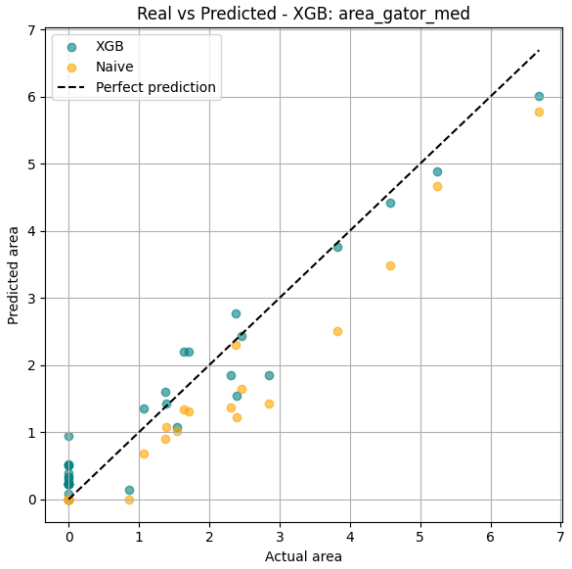



*Figure 47 Medium severity predictive vs real values for Naive and XGBoost models*

The **medium-severity** XGBoost model (Figure 47) presents a tighter clustering of points along the perfect prediction line, with a visible reduction in spread relative to the low-severity case. This model performs better than the Naive baseline resulting in more reliable and unbiased estimates throughout the full range of observed values and this is evident in the real areas between 1 and 5, where the Naive model tends to underpredict values. This improved correspondence demonstrates that the model successfully captures the intermediate deterioration behavior that depend on its own historical extent and high-severity cracking.

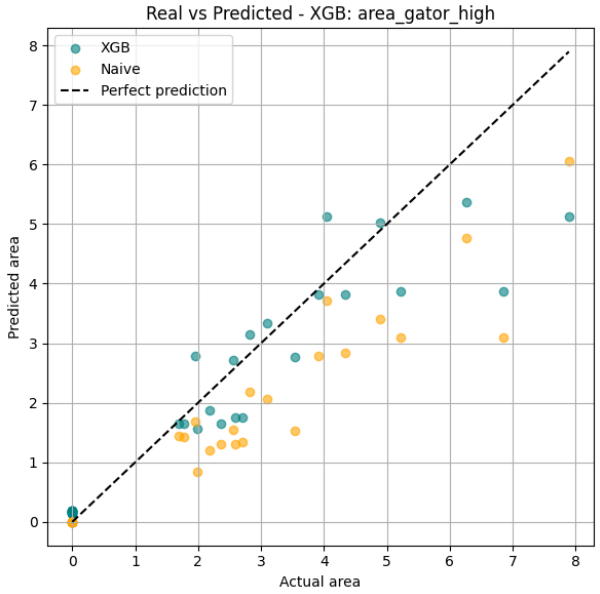



*Figure 48 High severity predictive vs real values for Naive and XGBoost models*

For **high-severity** deterioration and actual areas smaller than 5, the XGBoost approach definitely surpasses the Naive baseline, which underpredicts large areas of deterioration in a systematic way. The XGBoost results (see Figure 48), while a little more spread out, align more closely with the overall 1:1 trend, suggesting that there is an enhanced ability to model the nonlinear escalation patterns seen in more severe pavement distress. The variability at the highest levels of deterioration may also reflect the processes leading to and the inherent stochasticity behind severe cracking development, where localized structural failures and unmonitored external factors (e.g., drainage or other construction heterogeneity) influence the estimates.

- **Error Distribution Analysis**

The histograms show the distribution of prediction errors for each of the tuned XGBoost model and Naive baseline across the low, medium, and high severity levels (see Figure 49, Figure 50 and Figure 51).

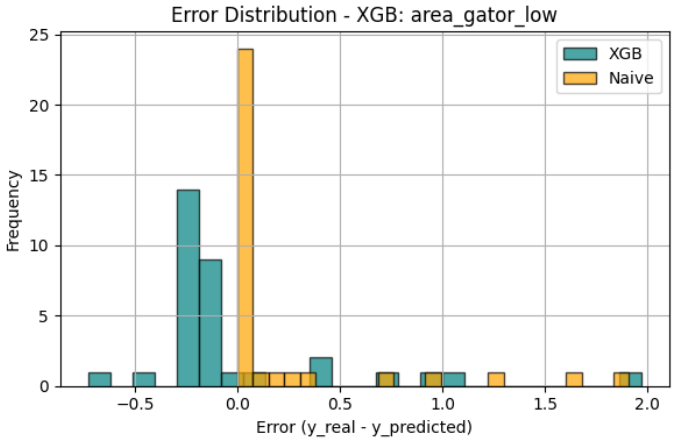



*Figure 49 Error distribution for the XGBoost model in low severity*

In the **low severity** subset (see Figure 49), the Naive model contains a distribution that is highly concentrated with a spike near zero with very little variance. That entire distribution is still offset from zero, confined to one bin located at approximately +0.1; exhibiting a persistent, systematic positive bias in which it tends to underestimate larger deterioration values, leading to occasional but significant deviations. Meanwhile, the XGB model shows a broader variance, encompassing the errors spanning from about -0.7 to 1.0. While this distribution of errors is broader, the maximum peak is located close to zero, exhibiting less bias overall and reflecting greater flexibility and generalization ability.

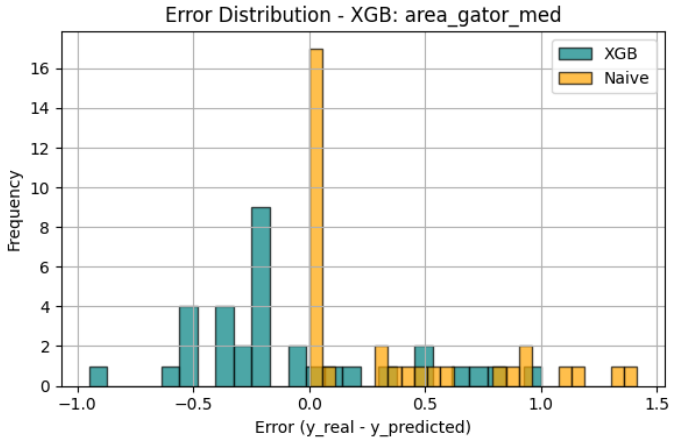



*Figure 50 Error distribution for the XGBoost model in medium severity*

For the **medium severity** evidenced in Figure 50, the error distributions for both models reveal distinct and imperfect behaviors. The XGBoost (XGB) model demonstrates a clear negative bias, with most error frequencies located between -0.75 and -0.25. This is a clear indication that the XGB model has a consistent tendency to over-predict (i.e., predicted values higher than real values leading to negative errors). The Naive model exhibits a positive bias, with the errors dispersed across the positive side of the histogram, indicating a tendency to under-predict. The Naive model does display a single, very high-frequency bin close to zero, suggesting that it accurately predicts value directions very often, but its errors are much more dispersed to the favorable side of the histogram (including a tail near approximately 1.4). Neither of the models shows a perfect center, but the XGB model's errors appear slightly more concentrated even though consistently negative.

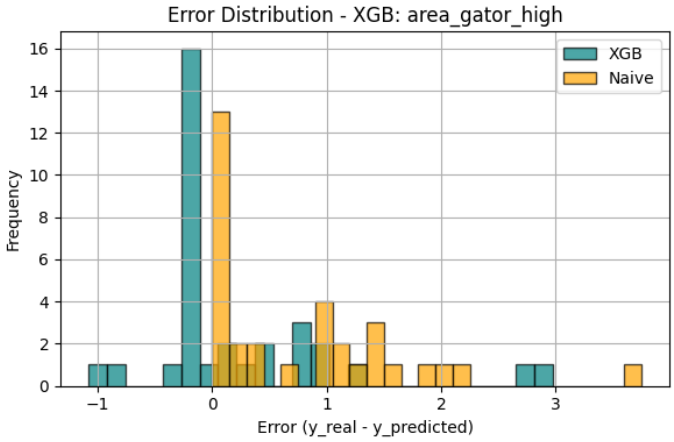



*Figure 51 Error distribution for the XGBoost model in high severity*

In the **high severity** case, the XGBoost model demonstrates a clear and significant performance advantage. The distribution of errors evidenced in Figure 51 is sharply peaked and quite tight around zero, as indicated by the tallest bar (about 16) located to the left of zero, representing an optimal error distribution, showing low bias (centered near zero) and low variance (narrow spread). This concentration also suggests that the XGB model is reliably accurate in predicting high-severity cases. In contrast, the Naive model's error distribution is flatter and much more dispersed, with a noticeable positive skew. This wide spread indicates high variance and a great deal more errors with large predictions particularly large under-predictions (errors > 1.0). Therefore, the XGB model is demonstrably more reliable and robust for predicting high-severity outcomes.

- **Learning Curve (MAE)**

Figure 52, Figure 53 and Figure 54 present the learning curve for low, medium and high severity XGBoost models.

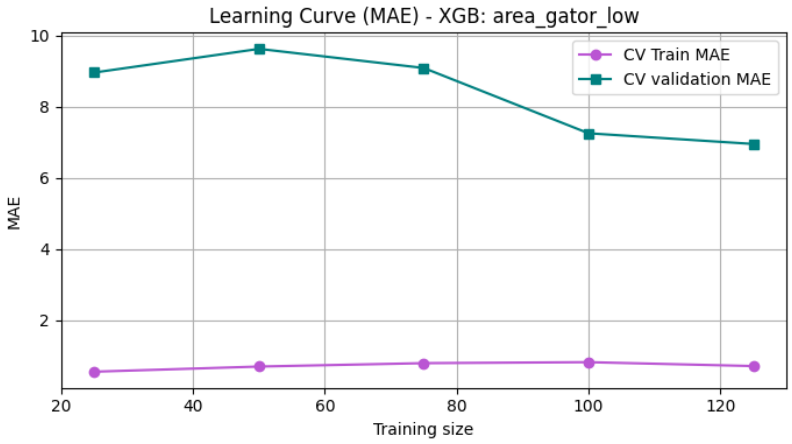



*Figure 52 XGBoost learning curve for low severity model*

The **low severity** model exhibits a very clear and extreme case of high variance (overfitting). The violet train MAE line illustrates this very clearly as it is flat and close to zero (ending at ≈0.3), meaning the model is perfectly "memorizing" the training data, and the green CV MAE line is extremely high (ending at ≈7), meaning the model has almost no ability to generalize this memorization to new, unseen data. This creates a huge "generalization gap" (real-world error is more than 10 times what the model thinks is its error). This diagnosis is backed up by the $R^2$ bar chart, showing the model thinks its perfect (Train $R^2 = 0.998$) while its real-world performance is much worse (Test $R^2 = 0.858$), only barely above the simple baseline.

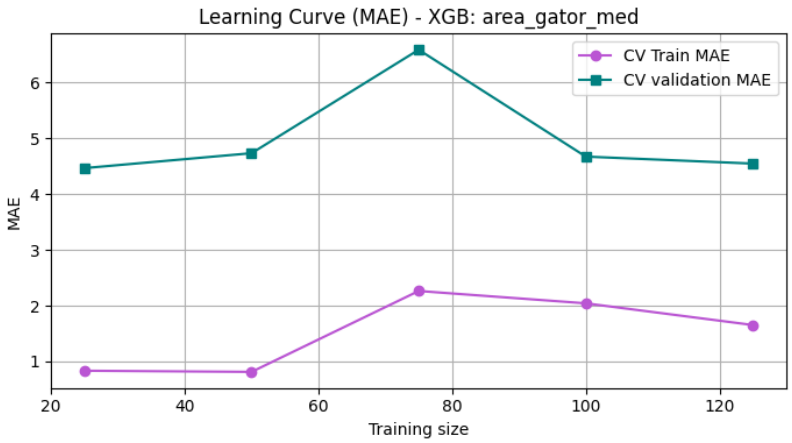



*Figure 53 XGBoost learning curve for medium severity model*

The **medium severity** model, as in the RF mode, depicts an unstable behave (see Figure 53). The training mean absolute error (MAE) is low across all training sizes (less than 2.5), and the validation MAE oscillates, with even large jumps at mid-range training sample sizes. The fluctuations in validation MAE indicate that performance is sensitive to the validation data composition, possibly due to considerable variability or influential outliers. In addition, the gap between training and validation MAE remains stable and considerable suggesting over-fitting, meaning the model fits well on the training data, but generalizing to unseen samples is difficult. This may occur due to medium severity observations not being very representative, or due to medium severity being heterogeneous and including mixed transitions from low to high severity observations, which adds yet another unstable and mixed signal. Adding more observations, using an additional level of regularization, or improving the representativeness of the features might alleviate some of these oscillations and improve the robustness of generalization.

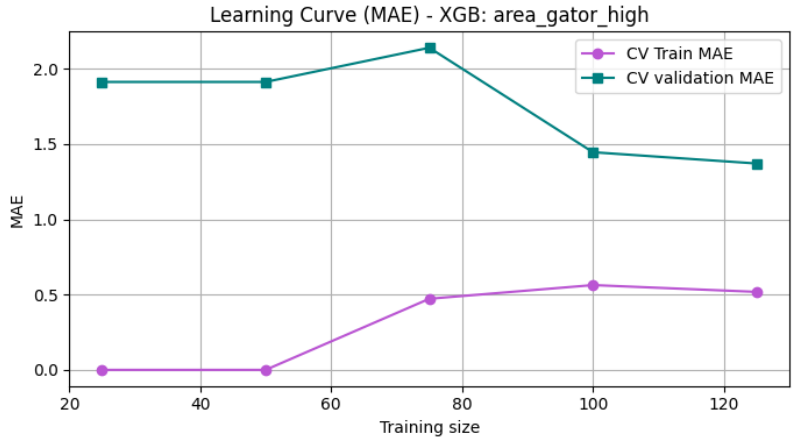



*Figure 54 XGBoost learning curve for high severity model*

On the other hand, the **high severity** model is the only model exhibiting a healthy, converging fit. The critical diagnostic in the plot is the small violet Train MAE line, which rises from 0 to 0.5. This is good because it indicates that the model is being constrained to stop "memorizing" and instead must learn a generalizable pattern. Simultaneously, the teal CV MAE line continues to decline with increasing data to around 1.4, again demonstrating its success in learning. Most notably, the two lines are converging, and the final generalization gap is small and reasonable (i.e., 0.5 train error vs 1.4 validation error). The dimensions of healthy learning come through effectively in the $R^2$ bar chart, which indicates the smallest and least acceptable performance gap between Train $R^2$ (0.909) and Test $R^2$ (0.850).

- **Discussion and Interpretation**

The analysis performed offers a thoughtful and meaningful examination of the XGBoost model, surpassing a mere claim of superiority, to assess its performance by severity. The main take away is that as in the RF model, the XGB model's main value is not in being an overall enhancement, but that it is a testing tool that precisely predicts high-severity outcomes. One important point of interest is that there is a notable difference between RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). The model clearly wins on the RMSE measure in all categories (particularly reducing RMSE by 24% in the "High" severity grouping) in indicating a strong ability to minimize large-cost prediction errors. Although the baseline model has a better (lower) MAE for low and medium severities, this illustrates that the XGBoost model captures a higher average error for less severe predictions, while preventing extreme errors.

The high severity model is clearly identified to be the best performing and most robust. It outperforms the baseline on all three critical metrics ($R^2$, RMSE, and MAE), confirms this numerical excess through decisive diagnostic tests. The distribution of errors is described as optimal due to its "peaked" and "well-clustered around zero" nature, enhancing it to low bias and low variance. Most importantly, the learning curve demonstrates that this is a model with a healthy, converging fit, suggesting that its strong performance is generalizable and not attributable to overfitted behavior. This is contrasted directly by the Naive baseline, which is demonstrated to be high variance and systematically under-predict critical failures.

In contrast, the models for low and medium severity levels are identified to have significant issues. The low severity model experiences overfitting, exhibiting a "huge generalization gap," or difference, between a near-perfect training $R^2$ of 0.998 and a fairly impressive test $R^2$ of 0.858, which is also substantiated by the corresponding learning curve in the model. The medium severity model is also characterized as "unstable" and has overfit, where error analysis suggests a somewhat consistent negative bias towards over-prediction. These elements indicate that the model architecture generally works, but it fails to generalize accurately under circumstances where variability is low (low severity) and / or where the data is a heterogeneous group of transition states (medium severity).

### 4.1.3 LightGBM Applying Transfer Learning (KNN)

- **LightGBM Hyperparameters**

  The process of Randomized Search with group K-Fold cross-validation, detailed in section 3.4.5, was used to determine the best-performing set of hyperparameters for each severity level for the lightGBM model. As Table 12 illustrates, the three severities share the same values for most of the hyperparameters.

*Table 12 Optimal LightGBM hyperparameters after tuning.*

| Hyperparameter | Low severity | Medium severity | High severity |
|---|---|---|---|
| **n_estimators** | 800 | 800 | 1200 |
| **max_depth** | 13 | 13 | 11 |
| **learning_rate** | 0.05 | 0.05 | 0.01 |
| **subsample** | 0.6 | 0.6 | 0.6 |
| **colsample_bytree** | 1.0 | 1.0 | 0.6 |
| **min_child_samples** | 5 | 5 | 5 |
| **reg_alpha** | 0.5 | 0.5 | 0.5 |
| **reg_lambda** | 1.0 | 1.0 | 0.5 |
| **num_leaves** | 15 | 15 | 95 |

- **Selection of K in KNN**

  Following the procedure detailed in Section 3.4.5, the metrics presented in Table 13 established k=10 as the final configuration. This selection was based on its strong performance in MAE value (MAE=1.244).

*Table 13 K values evaluated in lightGBM and its results*

| K value | MAE | R2 |
|---|---|---|
| **3** | 1.889 | -8.925 |
| **5** | 2.098 | -49.021 |
| **8** | 1.255 | -1.147 |
| **10** | **1.244** | **-1.437** |
| **15** | 1.373 | -3.584 |
| **20** | 2.353 | -5.423 |
| **30** | 1.353 | -3.831 |

- **Model Performance Overview**

The predictive performance results for LightGBM (LGBM) model across the three severity levels are shown in Figure 55, Figure 56 and Figure 57.

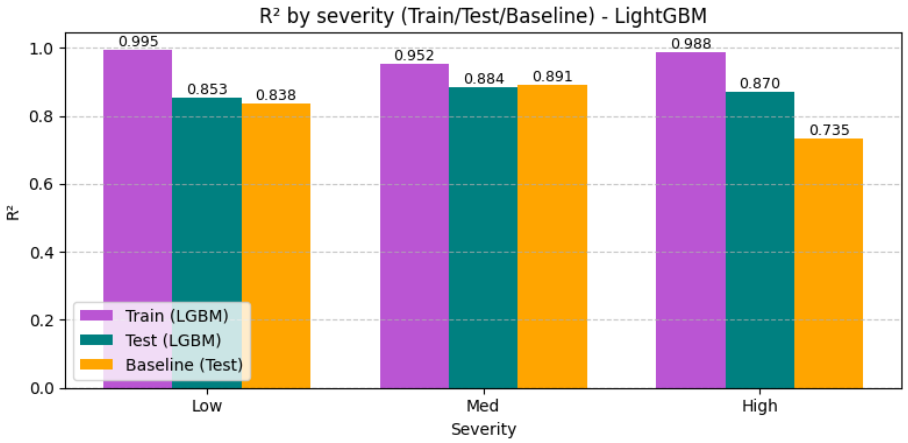



*Figure 55 LightGBM $R^2$ results by severity*

The LightGBM model has great predictive ability for all severity levels, but to different degrees of generalization (see Figure 55). In terms of goodness-of-fit, the model reports very high $R^2$ values on the training set (from 0.952 to 0.995), indicating very strong internal model fit. The test $R^2$ values ranging from 0.853 to 0.884, however, demonstrate that LightGBM does not generalize well with different degrees of generalization. For low severity, the model reports near perfect fit on the training set ($R^2$ = 0.995) and substantially lower performance on the testing set ($R^2$ = 0.853) which only slightly exceeds the baseline ($R^2$ = 0.838). This difference illustrates clear overfitting in the lower severity, further supported by the error analysis, where the model yields a higher MAE (0.31) than the baseline (0.23). The medium severity category results in a test $R^2$ value of 0.884 which is effectively similar to the baseline (0.891) and indicates there was no possible gain in capturing the variability despite a high training $R^2$ of 0.952. For high severity, LightGBM model is clearly outperforming the baseline the strong test $R^2$ =0.870 compared to the baseline's $R^2$ =0.735, combined with the substantial reduction in MAE and RMSE, suggests the model is generalizing effectively and capturing meaningful, complex patterns.

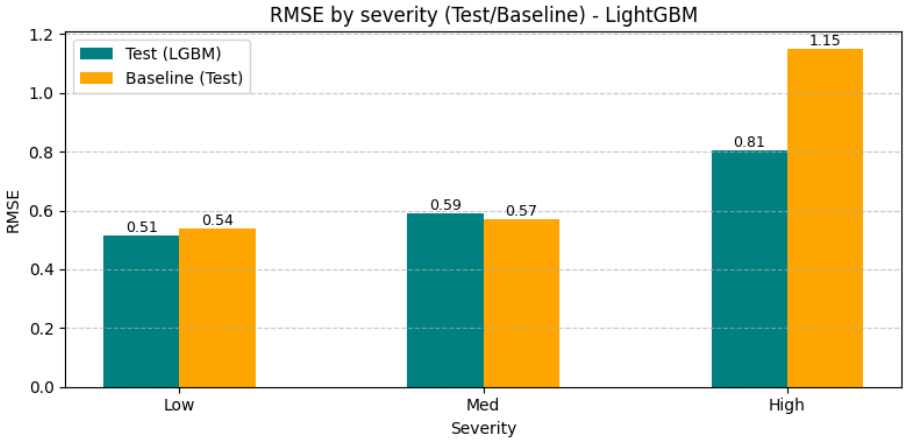



*Figure 56 LightGBM RMSE results by severity*

The LightGBM model shows a clear increase in predictive error with increasing severity level, a pattern also observed in the Baseline model (see Figure 56 and Figure 57). In the case of high severity level and for the LightGBM model, both MAE (0.57) and RMSE (0.81) are much larger than the errors for the low severity (MAE: 0.31, RMSE: 0.51) and medium severity (MAE: 0.38, RMSE: 0.59) levels, indicating that the high severity data is more difficult to predict, in general. In any severity category, the LightGBM model has mixed results relative to the Baseline. It performs significantly better than the Baseline at the high severity level (MAE: 0.57 vs 0.76, RMSE: 0.81 vs 1.15) which reinforces its performance with complex, highly variable data. However, in low severity, LightGBM has larger error than Baseline (MAE: 0.31 vs. 0.23) and mixed results at the medium severity (worse MAE, but better RMSE).

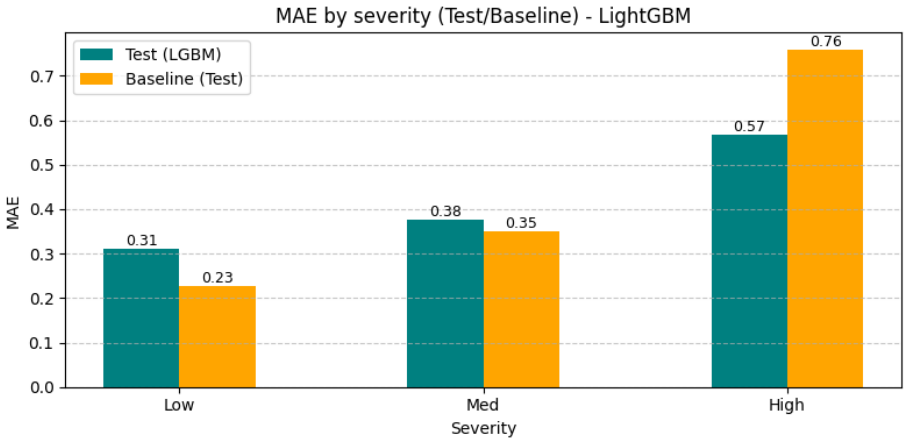



*Figure 57 LightGBM MAE results by severity*

- **Feature Importance Analysis**

An analysis of feature importance was conducted using the tuned LightGBM models to quantify the relative influence of each predictor on the estimated deteriorated area per severity class.

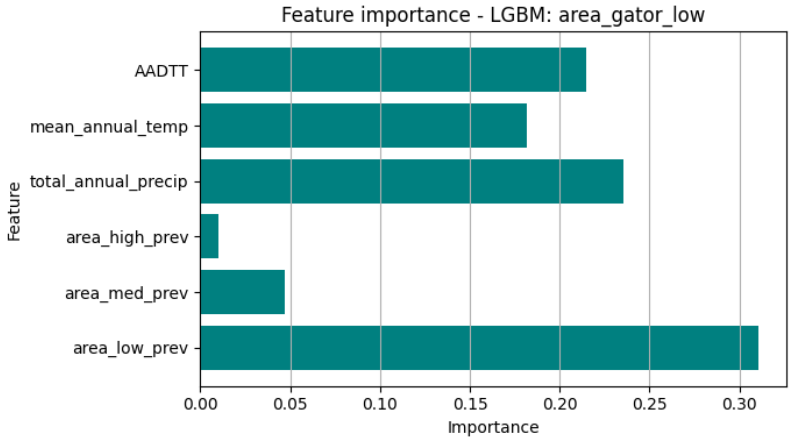



*Figure 58 LightGBM feature importance for low severity model*

As evidenced in Figure 58, to predict **low severity** areas, the model is heavily reliant on the historical low-severity area feature area_low_prev which takes the top position in the plot with an importance score greater than 0.30, indicating that the value of the previous low-severity area feature is the single largest predictor of low-severity for that current event. After area_low_prev, total_annual_precip (approx. 0.24) and AADTT (approx. 0.21) are heavily influenced in the model, again suggesting that climate and traffic features are the next most significant drivers, and mean_annual_temp is moderately significant as well (approx. 0.18).

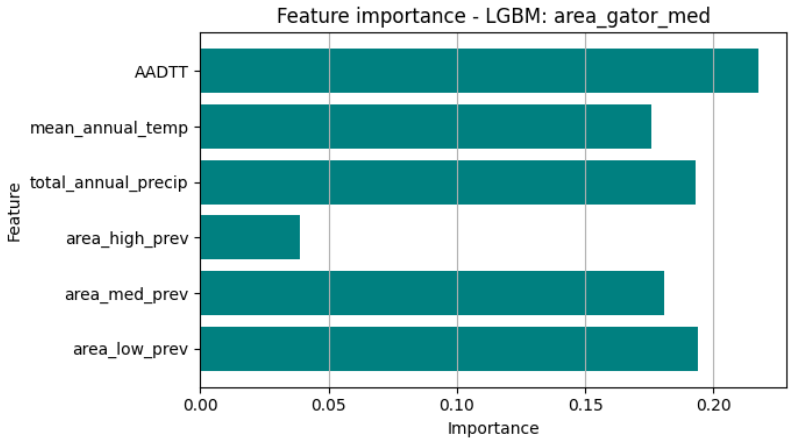



*Figure 59 LightGBM feature importance for medium severity model*

The feature importance profile for **medium severity** areas illustrated in Figure 59, shows a shift to a more balanced and distributed dependence on the various features. The predictive power is distributed much more evenly across the top four features, which are also better differentiated than in the low severity case, all with scores of approximately between 0.18 and 0.21. AADTT now emerges as the single most important feature (approx. 0.21), illustrating the pivotal role that traffic load plays on the occurrence of moderate severity areas. The other two features, total_annual_precip (approx. 0.20) and area_low_prev (approx. 0.19), also retain a high amount of importance indicating that cumulative moisture and even smaller scale historical events continue to have an appreciable influence. mean_annual_temp (approx. 0.17) also continues to retain a strong predictive role. As a point of interest, the historical features area_med_prev and area_high_prev were still of relatively low importance (approx. 0.17 and 0.04 respectively), although greater than in the low severity case. area_med_prev's importance increased compared to the low severity case, indicating the model is starting to pull from its own historical severity class, while other features remain more important.

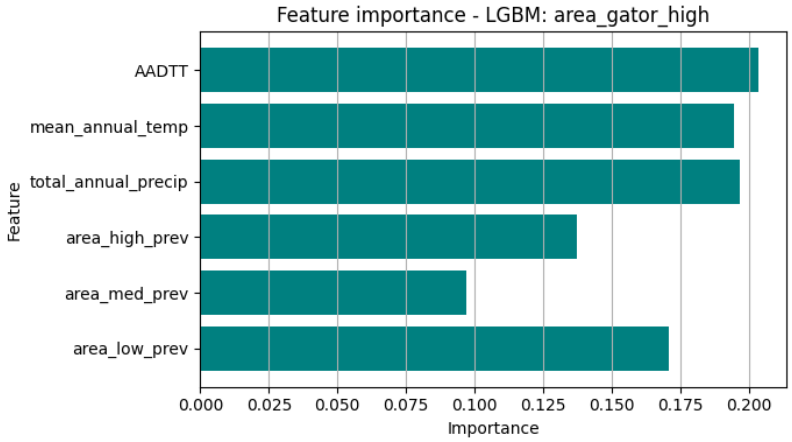



*Figure 60 LightGBM feature importance for high severity model*

The examination of the **high-severity** features demonstrates the most substantial shift in the model's predictive behavior (see Figure 60). All three of the top features (AADTT, mean_annual_temp, total_annual_precip) remain particularly important and are all evaluated at very close to 0.20 for importance, but their cumulative predominance was further lessened with the larger importance of the historical area features. Historical area features experienced general and substantial increases in importance about the other (non-historical) area features, and, in particular, area_high_prev is especially noteworthy for increasing significantly in

importance across all response categories, now being the fourth most important feature at an estimation of 0.14. The historical context provided multiple years of area high and area medium distant events that the model could use for prediction, and area_med_prev also had the importance rise to close to 0.10. This structural change hints that, for high severity, cumulative events are very important, and and event taken in context year before matters to predictive behavioral signals. This high severitty model is relatively the most complex because it incorporated traffic and climate data along with a layer of historical context of significantly larger previous events.

- **Predicted vs Observed comparison**

Figure 61, Figure 62 and Figure 63 illustrate the relationship between actual and predicted deteriorated areas for the LightGBM and the baseline (Naive) models.

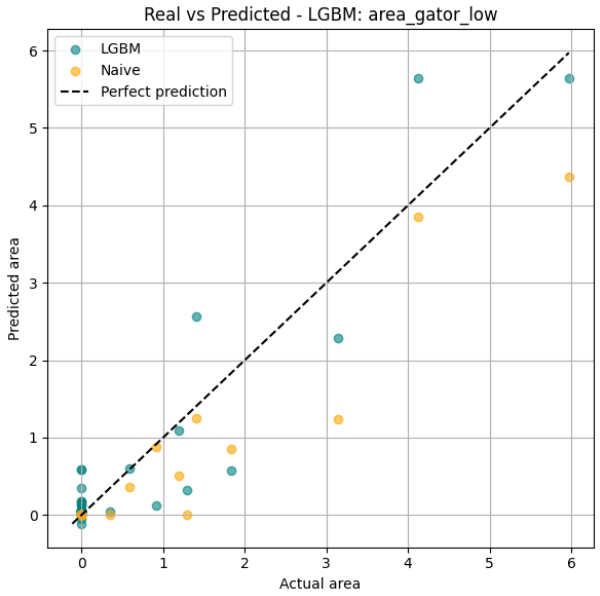



*Figure 61 Low severity predictive vs real values for Naive and LightGBM models*

With respect to the **low severity**, the results evidenced in Figure 61 show a widespread deviation from the optimal performance for LGBM, in particular, for larger actual areas. For areas below ≈ 1.5, both models show modest scatter, with the Naive model occasionally demonstrating predictions slightly closer to the line. Overall, the performance of LGBM in the low severity grouping is notably poor, presenting significantly problematic fidelity to the perfect line and over-predicting actual area at some points.

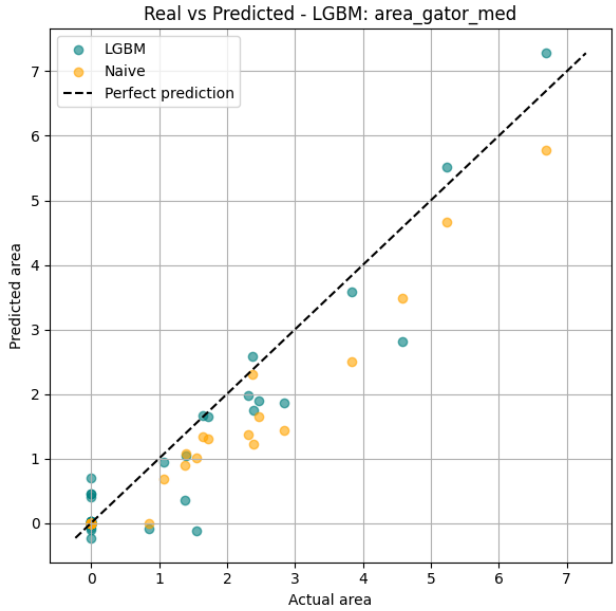



*Figure 62 Medium severity predictive vs real values for Naive and LightGBM models*

The **medium severity** outcomes are an appreciably improved performance characterized by a much tighter clumping of predicted points around the perfect prediction line for both models reflected in Figure 62 . Specifically, he LGBM model, exhibits greater ability to follow the upward trend in actual deterioration, especially for mid-range area values around 1–4. While there is still some under-prediction for LGBM around 4.5 and some over-predictions in greater area values, it is clear that the overall trend is showing that the features for the medium severity data are actually more informative, or less noisy and the LGBM model is therefore able to successfully model the information better than for the low-severity case.

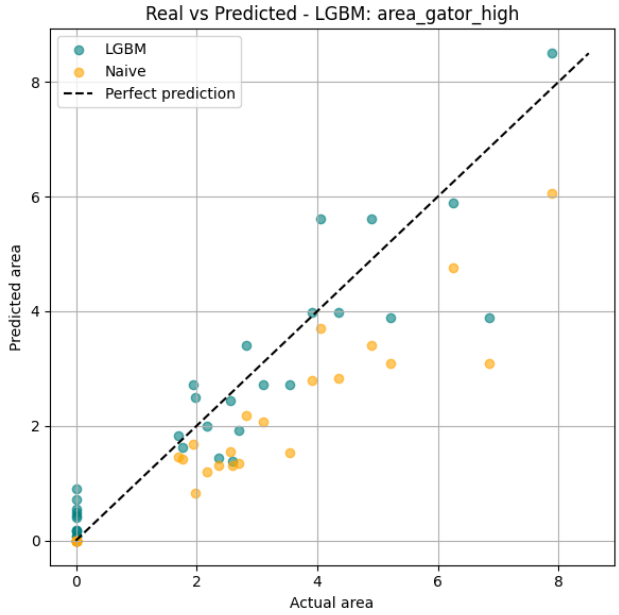



*Figure 63 High severity predictive vs real values for Naive and LightGBM models*

- **Error Distribution Analysis**

Figure 64, Figure 65 and Figure 66 present the distribution of prediction errors for both the LightGBM (LGBM) model and the Naive baseline, across the three severity levels by showing the frequency with which each model makes errors of different magnitudes.

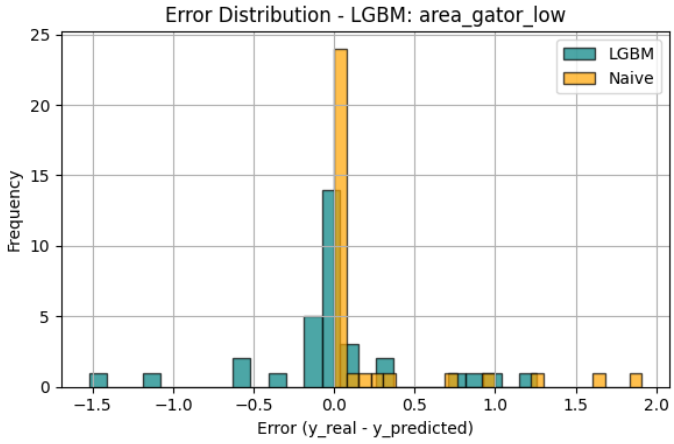



*Figure 64 Error distribution for the LightGBM model in low severity*

The error distribution for the **low severity** case exhibits a significant difference between the two models (see Figure 64). The LightGBM error distribution shows more spread of residuals than the naive model, most LGBM errors remain close to zero, while the model exhibits significant positive and negative deviation from zero, indicating frequent over- and underestimation of the true degree of deterioration.

On the other hand, the error distribution for the Naive model is highly concentrated around zero, indicating that a large count of predictions is very close to the actual value. It also indicates a notable asymmetry with a tail extending toward positive errors (1.5–2), indicating underestimation of actual deterioration in select instances.

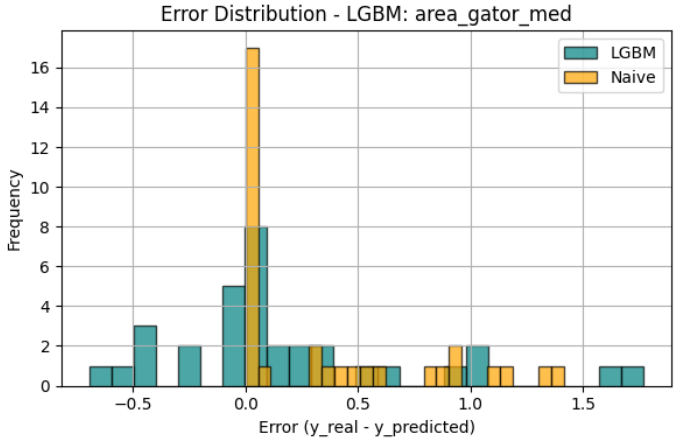



*Figure 65 Error distribution for the LightGBM model in medium severity*

In the **medium severity** case, the LightGBM residuals (see Figure 65) are more tightly distributed around zero than in the low severity case, indicating a more stable generalization. The distribution still has residuals in both directions, and several moderate underestimations and overestimations. The naive model is clustered around zero, with multiple positive errors which indicate it systematically underpredicts medium deterioration at larger true values. LightGBM captures more of the variability in real deterioration than the naive, but its spread of residual is still wide and consistent with the performance metrics which did not provide significant improvements over the baseline.

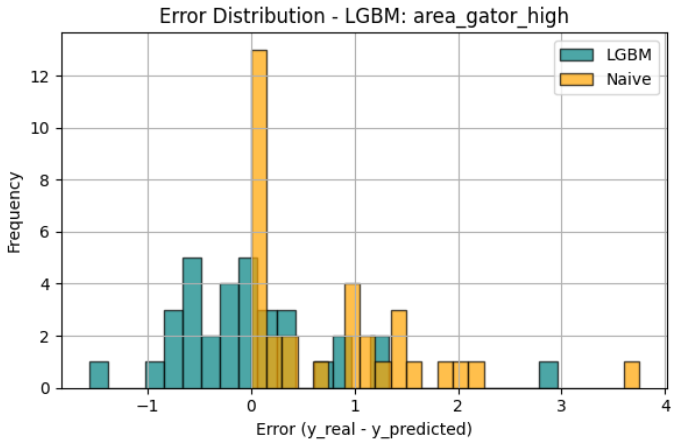



*Figure 66 Error distribution for the LightGBM model in high severity*

For **high severity**, the error distribution (see Figure 66) shows a broadening of the error spread for both models, which reflects increased prediction difficulty for extreme events. The LGBM error distribution is noticeably wider than medium severity, indicating overall higher variance with fewer extreme outliers demonstrating a relatively balanced near-zero mean error against the Naive baseline. Conversely, while the naive model has its highest frequency at zero error, it shows an evident bias towards positive errors which implies the model consistently underpredicts when the deterioration is severe.

- **Learning Curve (MAE)**

Figure 67, Figure 68 and Figure 69 present the learning curve for low, medium and high severity LightGBM (LGBM) models.

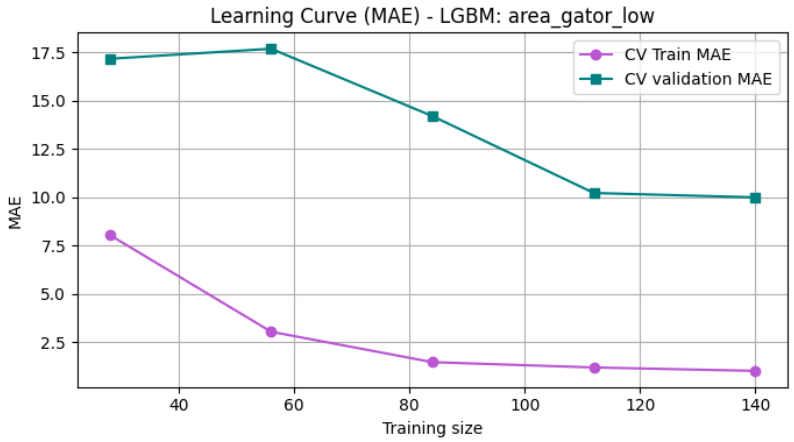



*Figure 67 LightGBM learning curve for low severity model*

The learning curve for **low-severity** deterioration evidenced in Figure 67, demonstrates a substantial and persistent gap between training and validation MAE, indicating strong overfitting. The CV Train MAE (purple) falls sharply with greater training size, finishing at a quite low MAE of about 1.0, which suggests that the LGBM model is able to both learn and memorize the training data very well. On the other hand, the CV Validation MAE (teal) begins at a really high level of about 17.5 and while it drops with more data, it eventually gets constant at a significantly high MAE of about 10.0. The large and persistent gap between training and validation curves is the definitive sign of overfitting. This pattern suggests the model has trouble generalizing to new data due to low variability and noise around early

stages of deterioration, behavior shown in the spread-out predictions depicted in the scatter plot above.

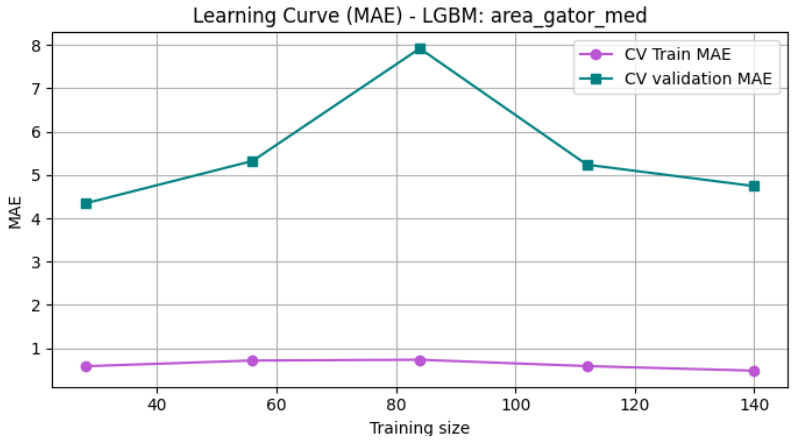



*Figure 68 LightGBM learning curve for medium severity model*

The **medium-severity** learning curve demonstrates a more stable but still imperfect generalization pattern. As shown in the Figure 68, the training MAE consistently remains low at all sample sizes (MAE ≈ 0.5-0.75), indicating a good within-sample fit. The validation MAE fluctuates considerably: it first increases and peaks at a MAE of 8.0 around the mid-range sample size of ≈ 85 then decreases again with larger samples. This shows that the model is moderately unstable and implies that the medium-severity deterioration has complex, mixed behavior that requires sufficiently large samples. The final validation MAE (≈ 4.7-4.9), while lower than the mid-range peak MAE, nonetheless remains considerably higher than the training error (again indicating there is some overfitting). More data appears to help with generalization, but the model still fails to completely capture the full variability associated with medium-severity propagation.

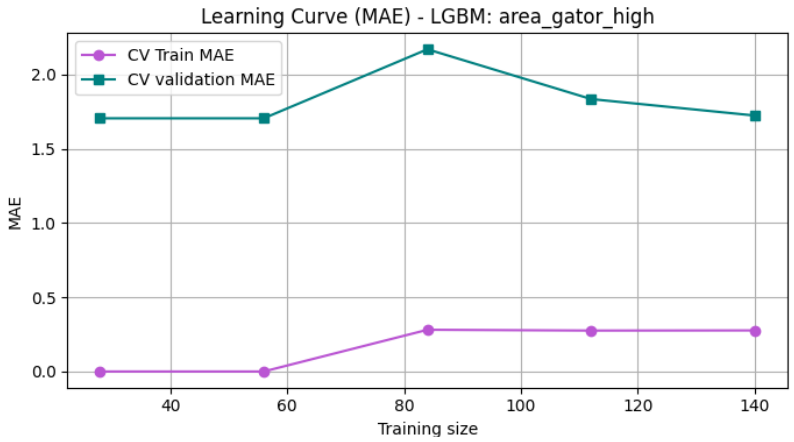



*Figure 69 LightGBM learning curve for high severity model*

The learning curve for **high-severity** yields the best generalization performance of all severities. An increase in sample size results in a slight increase in training MAE, indicating the model is not over-memorizing and is stabilizing. Validation MAE peaks moderately during the intermediate sample sizes but eventually decreases with larger training sets reaching values between 1.6 and 1.8. The gap between training and validation MAE remains the smallest across the three categories, showing much better stability of the model and low levels of overfitting. The fact that the training and validation curves are relatively flat suggests that simply adding more data of the same type may not, by itself, be sufficient to close this gap.

- **Discussion and Interpretation**

In essence, the analysis of the LightGBM model as in the case of RF and XGBoost has documented that its predictive performance is not homogeneous but rather, it is critically dependent upon the severity level of the target variable. The complexity of the model, a typical aspect of gradient boosting machines, is an advantage in the high-severity category of the target variable, which is defined by complex, long-term interactions. However, this same complexity has significant downsides in the low-severity category of the target variable, being especially subject to overfitting and performing worse than a simple baseline model.

The high-severity category is where the LGBM model illustrated its highest level of success. In this severity, the model "clearly outperformed" the baseline model, as evidenced by a higher test $R^2$ and, more significantly, the absolute values of MAE and RMSE were reduced. Additionally, the feature importance graphic indicates how the LGBM model achieved these success metrics, noting this is the "most complex" and diverse model. The LGBM model synthesizes data across various domains: climate, traffic (AADTT), and, of greatest importance, a deep historical aspect--specifically, in terms of the importance level of area_high_prev. The diagnostic plots confirm this, as the LGBM model corrects the bias of systemic under-prediction made by the baseline data--which produced a more balanced error distribution. Finally, the learning curve plot shows that this is the most generalizable model--in terms of the training and validation gap, which is the smallest, and an absolute validation MAE that is smallest (1.6-1.8).

In contrast, the low-severity category is a definitive failure case due to "strong overfitting," as indicated by the learning curve that shows that there is a "massive and

persistent gap" between a nearly perfect training MAE (1.0), and a very high validation MAE (10.0), leading to worse performance than baseline (MAE 0.31 vs. 0.23). The feature importance plot suggests that this is due to mechanistically simpler, "heavily reliant" on only the area_low_prev feature, which the model is ill-suited to capture with its complexity. The medium-severity case is an indifferent outcome; the model is "effectively similar" to the baseline and adds zero benefit ($R^2$ 0.884 vs. 0.891). Its learning curve is "moderately unstable"; it did not experience the severe failure that the low-severity model showed, yet it also did not achieve the performance generalizability that the high-severity model demonstrated.

In conclusion, the LGBM model is not a universally preferred approach; it will only be valuable for a complex enough problem. The LGBM model has value in modeling the high-severity events because, as the feature importance analysis highlights, there are multiple interacting historical, climate, and traffic factors captured in the layer and time history of the input, ultimately capturing or modeling complexity. For the more mechanistically simple low- and medium-severity events, adding this complexity only serves to worsen errors in the baseline solution.

## 4.2 Key Findings

This section compiles the main findings from the comparative assessment of the Naive, Random Forest (RF), XGBoost (XGB), and LightGBM (LGBM) models. As demonstrated in the previous model-specific evaluations, predictive performance is not consistent across severity levels and is strongly reliant on the particular deterioration category being modeled, in this way, this chapter first examines the model that performs best within each severity category and then summarizes the overall comparative performance across all severities.

### 4.2.1 Best-Performing Model by Severity Level

The next section selects the best model for each severity class, based on the comparisons of accuracy, generalization ability, and error behaviors evidenced in Table 14.

*Table 14 Model performance summary*

| Severity | Baseline | | | RF +KNN | | | XGBoost + KNN | | | LightGBM + KNN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE | $R^2$ |
| **Low** | 0.54 | 0.23 | 0.84 | 0.47 | 0.26 | 0.88 | 0.5 | 0.34 | 0.86 | 0.51 | 0.31 | 0.85 |
| **Medium** | 0.57 | 0.35 | 0.89 | 0.67 | 0.28 | 0.85 | 0.45 | 0.38 | 0.93 | 0.59 | 0.38 | 0.88 |
| **High** | 1.15 | 0.76 | 0.74 | 0.73 | 0.42 | 0.89 | 0.87 | 0.54 | 0.85 | 0.81 | 0.57 | 0.87 |

- **Low Severity Winner: Naive Baseline**

  This regime is strongly dominated by the persistence dynamics of the deterioration. The baseline already has a very low MAE and a reasonable $R^2$ leaving little room for improvement. RF evidently helps increase the goodness-of-fit $R^2$ and reduces RMSE, versus the baseline at the expense of slightly worse MAE but it also shows a learning curve that reveals a clear sign of overfitting. In the XGBoost and LightGBM models there are higher overfitting, showing near-perfect training fits but marginal gains or even performance degradation on test errors. This suggests that adding the complexity of gradient-boosted models does not help and, in fact, works against the performance here.

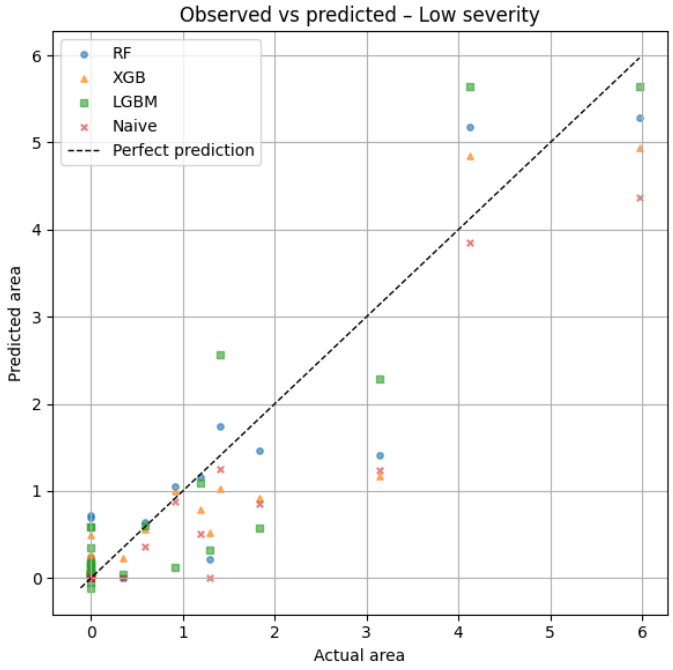



*Figure 70 Low severity predictive vs real values for all the models*

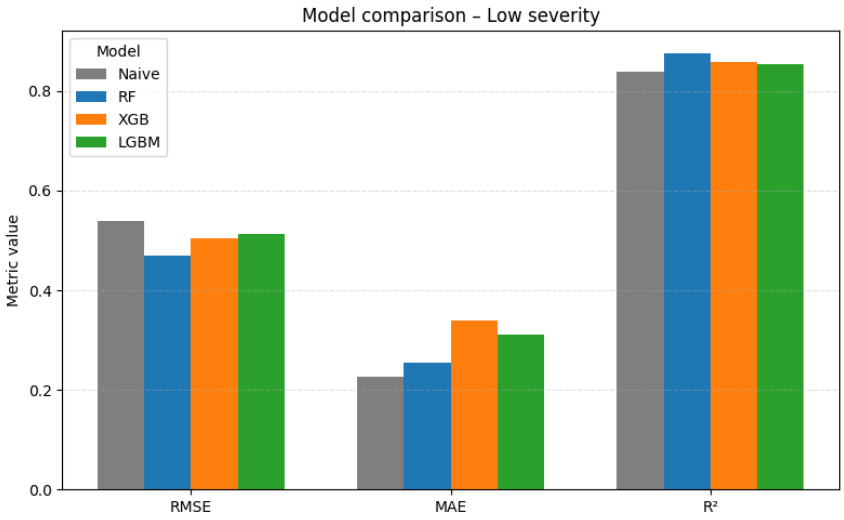



*Figure 71 Low severity predictive performance metrics for all models*

- **Medium Severity Winner: XGBoost (XGB)**

  This constitutes the most complex and troublesome predictive situation. The evolution of the deterioration is determined by both the persistence of moderate severity and the transitions from high severity, leading to richer non-linear interactions between the variables. In these terms, XGBoost is the clear winner, providing the best test $R^2$ and the lowest RMSE overall. This means it is clearly better at supporting extreme deviations (even if its MAE is slightly worse than the baseline). On the other hand, the predictions of the Random Forest indicated a smaller MAE than the baseline, but a lower $R^2$ and RMSE, suggesting it was likely being penalized by outliers in the squared error. Ultimately, LightGBM's prediction was very similar to the baseline, which meant there did not seem to be any improvements.

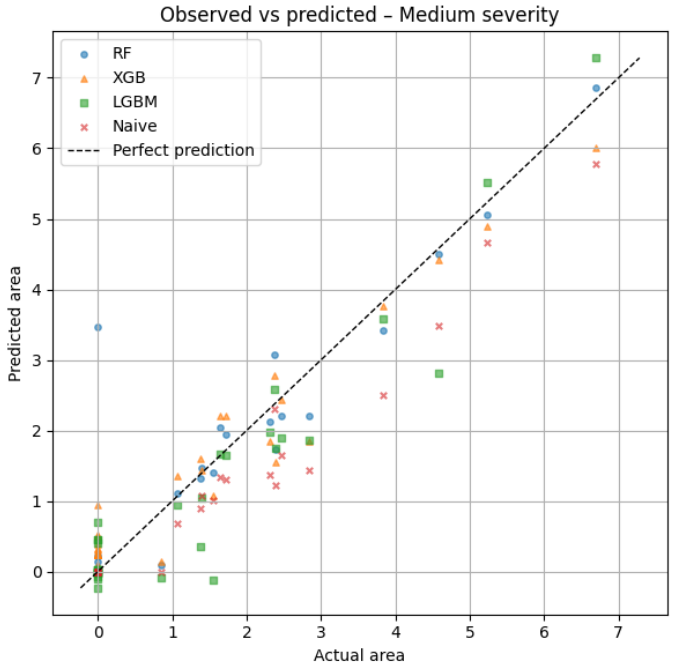



*Figure 72 Medium severity predictive vs real values for all the models*

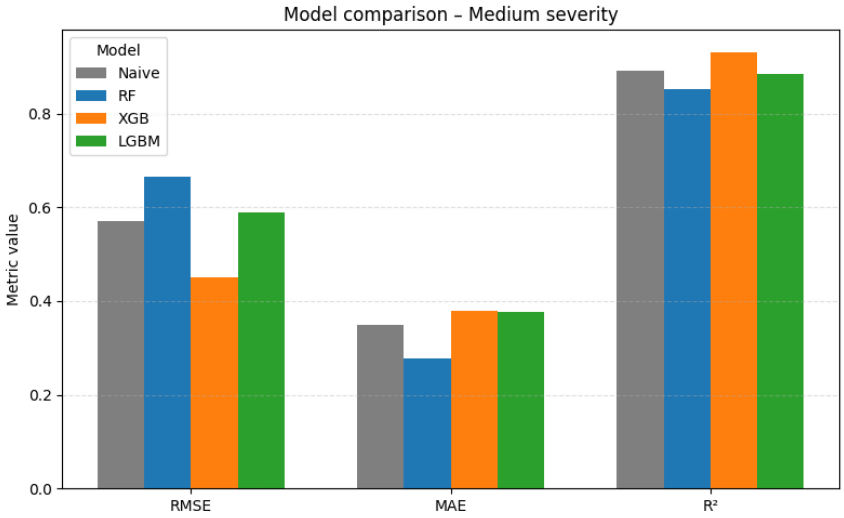



*Figure 73 Medium severity predictive performance metrics for all models*

- **High Severity Winner: Random Forest (RF)**

  In this category, the variability in performance is sharper and clearly benefit the Random Forest approach. The RF model produces the largest test $R^2$ values, the smallest MAE and RMSE values, all above the baseline, XGBoost, and LightGBM. Additionally, the error distribution is nearly symmetric about zero, and learning curves show a pattern of improvement in a stable pattern, indicating solid generalization capability. XGBoost also

clearly improves on the baseline, has healthy learning curve behavior, but is still systematically below RF. LightGBM though an improvement on the baseline, lags behind the other two models. Therefore, RF with KNN is the preferable operational option for high-severity predictions.

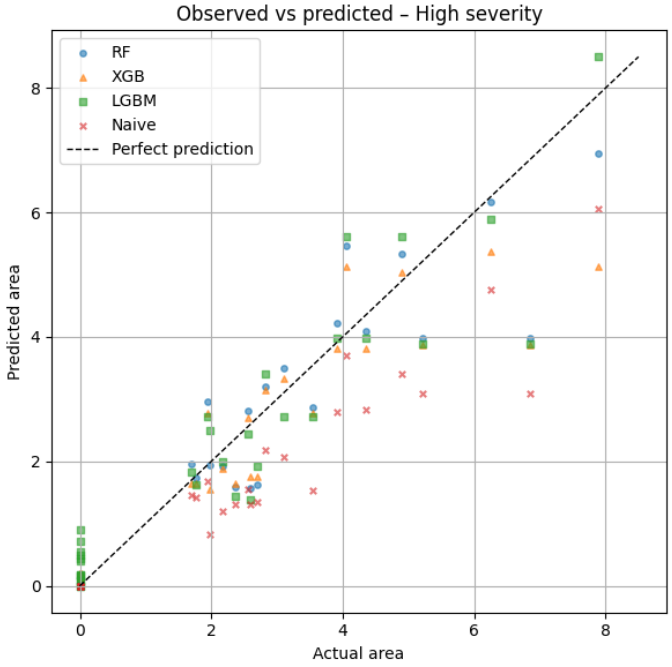



*Figure 74 High severity predictive vs real values for all the models*

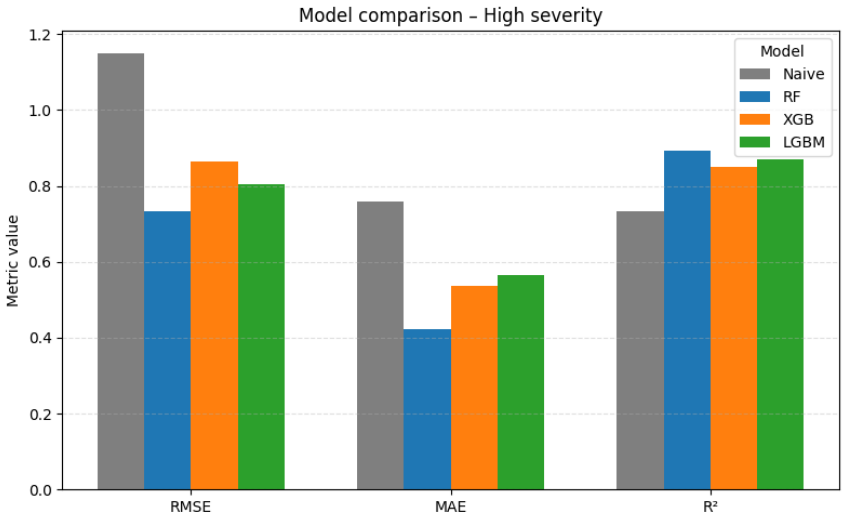



*Figure 75 High severity predictive performance metrics for all models*

### 4.2.2 Overall Model Efficacy and Comparison

Based on a global analysis of the results, the Random Forest (RF) model with the Transfer Learning (KNN) method stands out as the strongest and the most balanced method of those tested in terms of overall severity prediction for deteriorating areas.

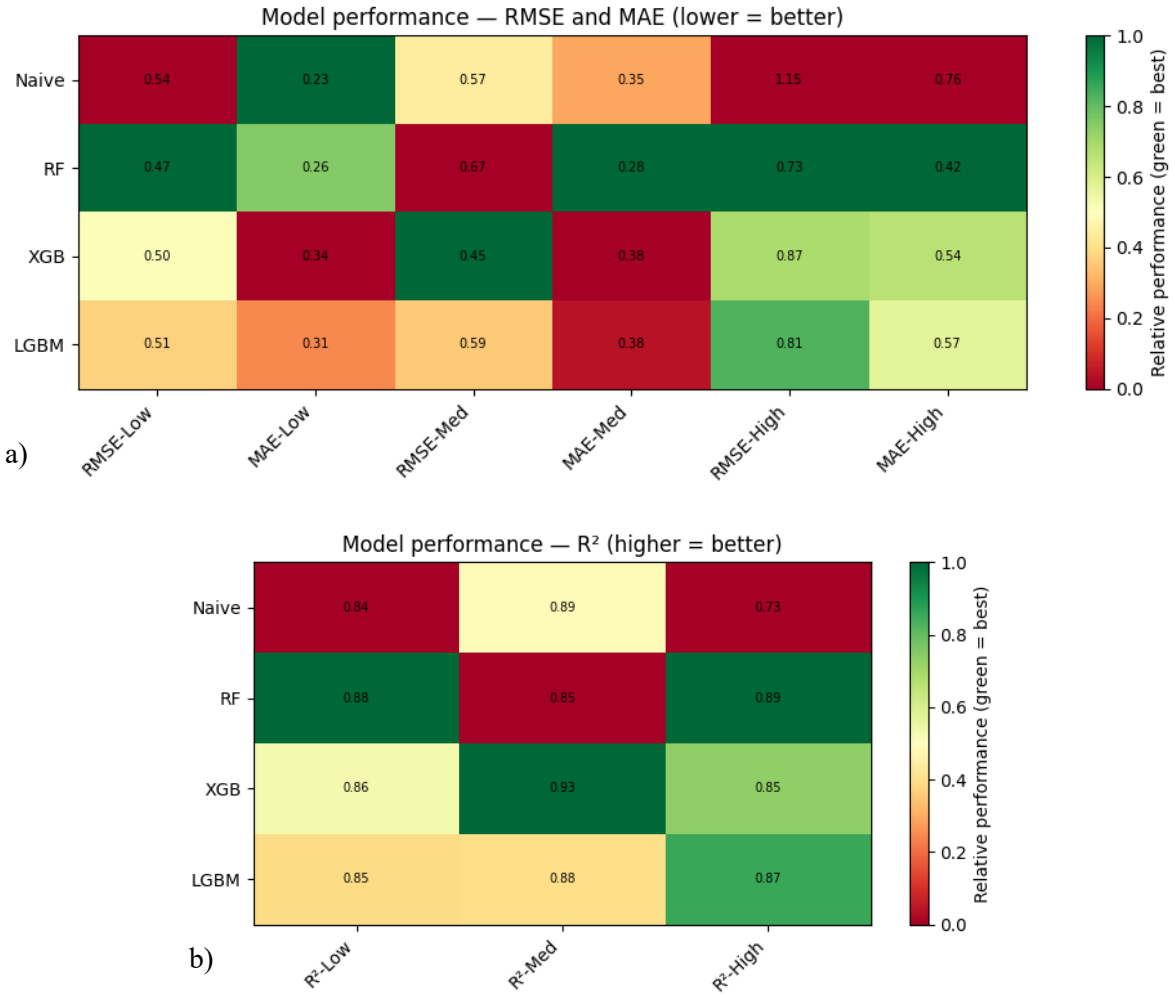



*Figure 76 Heatmaps of standardized error and accuracy profiles for all the models a) Heatmap of RMSE and MAE b) Heatmap of $R^2$*

This overall conclusion does not suggest uniformly superior performance in every circumstance. The comparative heatmaps (Figure 76) demonstrate that no model was completely superior to the other model; indeed, every model had its own set of strengths and weaknesses. The RF model demonstrated significant weaknesses in some conditions, specifically, it displayed a clear propensity to overfit in cases of low severity, where its complexity governed it to induce no value to the simple baseline model. In addition, it exhibited notable instability in the medium severity regime, where it suffered from an outlier

burden, achieving the worst RMSE performance of all the methods tested (-10% degradation vs. baseline).

*Table 15 Percentage improvement vs baseline. The positive sign (+) indicates an improvement, not an increase in the value*

| Severity | RF +KNN | | | XGBoost + KNN | | | LightGBM + KNN | | |
|---|---|---|---|---|---|---|---|---|---|
| | ΔRMSE | ΔMAE | $\Delta R^2$ | ΔRMSE | ΔMAE | $\Delta R^2$ | ΔRMSE | ΔMAE | $\Delta R^2$ |
| **Low** | +7% | -3% | +4% | +4% | -11% | +2% | +3% | -8% | +1% |
| **Medium** | -10% | +7% | -4% | +12% | -3% | +4% | -2% | -3% | -1% |
| **High** | +42% | +34% | +15% | +28% | +22% | +11% | +34% | +19% | +13% |

Nonetheless, the RF's overall superiority is based on its clearly excellent ability to perform on the most critical and complex task: predicting high-severity deterioration. In this task, RF showed strong performance superiority versus the other models. As shown in the percentage improvement table (Table 15), RF showed the largest performance advantage in this task with a 42% RMSE improvement and a 34% MAE improvement relative to the baseline. More importantly, the error analysis confirmed that RF was able to correct the baseline's systematic under-prediction bias, which is the most serious error from an infrastructure management perspective.

Therefore, while RF was not the most robust model in all circumstances, it is perhaps the most overall robust because it better handles the most important and toughest aspect of the prediction problem. RF can be viewed as the most reliable prediction tool for forecasting critical structural failures.

# 5

# CONCLUSIONS AND FUTURE WORK

This chapter synthesizes the main insights gained from the comparative modeling study and considers their implications for pavement management practice. It also identifies several limitations—particularly limitations in temporal depth, feature availability, and regional variability—that compromise the robustness and generalizability of the findings. Based on these limitations, a set of recommendations for improvements and future research directions are made, including enhanced data collection, advanced feature engineering, cross-regional validation, and next generation frameworks integrating probabilistic modeling and automated visual prediction.

## 5.1 Practical Implications for Pavement Management

The findings of this study have direct uses for pavement management systems, especially in contexts where historical data are limited and deterioration happens at different rates across severity levels. Although three machine learning methods were tested, the Random Forest (RF) model was the most useful. It was especially good at predicting high-severity alligator cracking—the point where deterioration becomes a big structural and financial problem for road agencies.

### 5.1.1 Improved identification of high-risk segments

RF gives much more reliable predictions in the high-severity range, lowering both average error (MAE) and large-error events (RMSE). In practice, this helps agencies to:

- Find sections at risk of structural failure sooner
- Prioritize timely repairs before expensive, full-depth repairs are needed
- Reduce unplanned maintenance and the disruptions it causes

This is very important because high-severity cracking is the most obvious sign of structural fatigue and imminent failure. Being able to predict this failure is key for risk mitigation and road safety, as it allows agencies to repair hazardous conditions before they cause accidents or vehicle damage.

### 5.1.2 More efficient allocation of Maintenance Budgets

Since keeping up pavement costs a lot (like ANAS's €1.6 billion yearly program), even small gains in prediction accuracy can lead to big savings. By distinguishing which sections will likely get worse within a year, RF supports:

- Targeted interventions
- Optimized scheduling of maintenance teams,
- Less reactive/corrective maintenance,
- Budgeting for upcoming years based on predicted deterioration trends.

### 5.1.3 Stronger Performance Despite Environmental Changes

RF captures the interaction between historical cracking patterns and climatic stressors like precipitation and temperature fluctuations without becoming unstable or overfitted. This lets agencies:

- Predict where environmental conditions will cause faster cracking,
- Support maintenance plans that respond to the climate,
- Add climate trends into pavement management plans.

### 5.1.4 Applicability in data-scarce contexts

Because the Italian data only has two time-based observations per section, standard deterioration models can't be reliably set up. The RF + KNN transfer learning system takes

care of this by using similar structural patterns from the U.S. LTPP data. This has important results:

- Cities with minimal monitoring can still get reliable forecasts,
- Not having enough data is no longer a reason not to use predictive maintenance tools,
- Agencies can slowly improve the model as they get more local data.

### 5.1.5 Less Uncertainty in Severity Classification and Inspection

As noted in the methods and research review, it's easy to make mistakes when labeling distress severity due to personal opinions and image-based problems. RF mitigates these problems by:

- Averaging results from multiple sources,
- Reducing the impact of wrongly labeled or flawed samples,
- Making more consistent year-to-year predictions.

This results in a more dependable decision-support system, even when inspection data isn't perfect.

### 5.1.6 Supports Proactive and Preventive Maintenance Policies

Most importantly, the model works best in areas where engineering principles predict the biggest structural and safety consequences: high-severity cracking. Because of this, RF allows agencies to:

- Move from reactive to preventive maintenance,
- Fix issues before they reach critical levels,
- Lower risk for road users by spotting structural failures early,
- Add predictive analytics to current Pavement Management Systems (PMS).

## 5.2 Suggestions for Improvement

Some limitations were identified in the present study which indicate obvious avenues for improvement. Enhancing temporal depth, incorporating additional structural and material characteristics, and increasing cross-regional validation would support improvements in model accuracy and transferability. The following subsections detail these important areas for improvement.

### 5.2.1 Availability of data and temporal depth

Only two temporal observations exist per road segment in the Italian dataset (t=2), thereby limiting the model's possibility of learning the temporal dynamics of deterioration. This lack of data increases the potential of overfitting models and restricts the opportunity of the data-hungry models like Random Forests. More importantly, the short temporal dimensions prevent the utilization of more complex sequential deep-learning models such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) that were developed to represent long-term dependency and temporal change in time series data. Therefore, an expansion of the temporal dimensions for the Italian dataset either through continuous monitoring, or a complementary set of data from another national sample, will be essential. Such enhancement would not only strengthen the robustness of current models but also enable the exploration of these deep learning architectures to evaluate their comparative performance in predicting pavement deterioration trends.

### 5.2.2 Enhanced feature engineering

The present study relies on a limited set of climatic and traffic indicators. However, there are other factors that influence the structural deterioration as well such as pavement layer thickness, material type, subgrade attributes and climate interactions with seasonality. Although comprehensive datasets such as the Long-Term Pavement Performance (LTPP) program often contain this type of structural and materials related information, in most practical pavement management databases that exist this type of information is not found.

By including these types of features, would be possible to provide models with more physically meaningful predictors, and yield an overall more reliable predictive performance which might reduce the predominant reliance on last year's distress level as the predictor. Therefore, when considering data collection in the future, efforts should prioritize the collection of structural and materials data to develop predictive models that are reliable and transferable.

### 5.2.3 Validation and Cross-Regional Generalizability

The entire training and validation of the models was performed within the Italian network, and the models shown a great generalizability but construction standards, materials,

environmental conditions, and maintenance procedures differ considerably by country. Future improvements should include:

- External validation using datasets from other countries.
- Cross-country harmonization of the distress definitions, measurement procedures, and attributes.
- Domain adaptation methods (in addition to KNN), such as TrAdaBoostR2, CORAL or adversarial adaptation to enable improved transfer of models across heterogenous networks.

Such improvements would help ensure that deterioration models can be implemented in operational settings across different pavement management systems.

## 5.3 Future Directions

In addition to immediate methodological improvements, promising research pathways emerge from the study's limitations and findings. These pathways are aimed at expanding on the predictive framework, improving its operational utility, and applying next generation data collection and modeling methods.

### 5.3.1 Probabilistic Modeling of Severity Transitions

The existing models provide point estimates for the area of deterioration, but do not provide any measure of the likelihood of a pavement segment transitioning from one severity class to another, e.g., what is the probability that a segment in low severity would transition to medium severity in one year, or what is the probability that pavement that is considered medium severity would transition to high severity in a two-year horizon? Estimating these transition probabilities would give pavement managers an improved basis for risk-based planning and maintenance prioritization.

Changing from deterministic predictions to the use of probabilistic transition modeling would allow agencies to identify segments that are at high risk of rapid deterioration and to prioritize maintenance not only on forecasted outcomes, but also on probabilities of severe declines in the future trajectories.

### 5.3.2 Integrated Automated Visual Detection and Predictive Forecasting

A promising future direction involves developing an end-to-end automated system of combining visual distress detection using deep-learning and forecasting deterioration one year ahead (or multiyear, depends on data). Given the recent advancements in computer vision, it is now feasible to recognize and characterize cracking patterns, rutting, potholes, and other surface distresses in real-time from images taken from onboard cameras in vehicles or drones. The combination of such models into a deterioration framework would produce a two-step pipeline.

- **Automated Distress Detection:** A deep-learning visual model with the capacity to detect and quantify distress areas directly from imagery, yielding objective high-resolution measurements that can supplement, or even supplant, manual inspection data. This significantly reduces the uncertainty that may stem from human subjectivity and variability in camera angles and survey intervals.
- **Short-Term Prediction of Future Deterioration:** The automatically-acquired distress metrics would be used to construct a prediction model—e.g., using Random Forest, LSTM, or a multi-task learning architecture—that would provide a prediction of the expected deterioration next year (e.g., next year distress area or severity class transitions).

The integration of models would enable a pavement management organization to automatically identify current distress conditions from image data, update the database in near–real time, and construct one-year-ahead predictions (no human interaction required).

# REFERENCES


[1] Federal Highway Administration (FHWA), "Long-Term Pavement Performance (LTPP) InfoPave™ Database," 2023.

[2] "LOKI srl." Accessed: Oct. 13, 2025. [Online]. Available: https://www.lokisrl.eu/startup/

[3] R. Haas, W. R. Hudson, and J. P. Zaniewski, *Modern Pavement Management*. Krieger Publishing, 1994.

[4] AASHTO (American Association of State Highway and Transportation Officials), *AASHTO Guide for Design of Pavement Structures*. AASHTO, 1993.

[5] ARA Inc. and ERES Division, "Guide for Mechanistic-Empirical Design of New and Rehabilitated Pavement Structures," 2000.

[6] B. Karki, S. Prova, M. Isied, and M. Souliman, "Neural Network Approach for Fatigue Crack Prediction in Asphalt Pavements Using Falling Weight Deflectometer Data," *Applied Sciences (Switzerland)*, vol. 15, no. 7, Apr. 2025, doi: 10.3390/app15073799.

[7] M. T. Tiza, "An Appraisal of Mechanistic-Empirical Models (MEMs) in Pavement Deterioration," *Journal of Nature, Science & Technology*, vol. 3, no. 2, pp. 1–10, Apr. 2023, doi: 10.36937/janset.2023.6855.

[8] A. Hu, Q. Bai, L. Chen, S. Meng, Q. Li, and Z. Xu, "A review on empirical methods of pavement performance modeling," Aug. 01, 2022, *Elsevier Ltd*. doi: 10.1016/j.conbuildmat.2022.127968.

[9] T. Tamagusko, M. Gomes Correia, and A. Ferreira, "Machine Learning Applications in Road Pavement Management: A Review, Challenges and Future Directions," Dec. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/infrastructures9120213.

[10] N. Sholevar, A. Golroo, and S. R. Esfahani, "Machine learning techniques for pavement condition evaluation," Apr. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.autcon.2022.104190.

[11] P. Marcelino, M. de Lurdes Antunes, E. Fortunato, and M. C. Gomes, "Transfer learning for pavement performance prediction," *International Journal of Pavement Research and Technology*, vol. 13, no. 2, pp. 154–167, Mar. 2020, doi: 10.1007/s42947-019-0096-z.

[12] J. Li, J. Guo, B. Li, and L. Meng, "Novel Instance-Based Transfer Learning for Asphalt Pavement Performance Prediction," *Buildings*, vol. 14, no. 3, Mar. 2024, doi: 10.3390/buildings14030852.

[13] A. Hemed, L. Ouadif, L. Bahi, and A. Lahmili, "Impact of climate change on pavements," 2020, doi: 10.1051/e3sconf/20.

[14] I. M. Ud Din, M. S. Mir, and M. A. Farooq, "Effect of Freeze-Thaw Cycles on the Properties of Asphalt Pavements in Cold Regions: A Review," in *Transportation Research Procedia*, Elsevier B.V., 2020, pp. 3634–3641. doi: 10.1016/j.trpro.2020.08.087.

[15] D. Levinson and K. Krizek, *The End of Traffic and the Future of Transport.* NETWORK DESIGN LAB, 2017.

[16] ANAS - Struttura territoriale Piemonte e Valle d'Aosta, "Dati Finanziari 2024." Accessed: Sep. 22, 2025. [Online]. Available: www.stradeanas.it/it/lazienda/dati-finanziari

[17] CEIC, "Road Infrastructure Maintenance: Constant Euro from 1995 to 2021." Accessed: Sep. 22, 2025. [Online]. Available: https://www.ceicdata.com/en/italy/transport-infrastructure-investment-and-maintenance-oecd-member-annual/it-road-infrastructure-maintenance-constant-euro

[18] European Commission, "Road safety thematic report – Main factors causing fatal crashes," Brussels, Apr. 2024.

[19] OECD/ITF, "Asset Management for Sustainable Road Funding," Paris, May 2013. Accessed: Sep. 22, 2025. [Online]. Available: https://www.oecd.org/en/publications/asset-management-for-sustainable-road-funding_5k46l8wh9lhg-en.html

[20] L. Titus-Glover, M. I. Darter, and H. Von Quintus, "Impact of Environmental Factors on Pavement Performance in the Absence of Heavy Loads," 2019.

[21] M. S. Rahman and S. Erlingsson, "Predicting permanent deformation behaviour of unbound granular materials," *International Journal of Pavement Engineering*, vol. 16, no. 7, pp. 587–601, Aug. 2015, doi: 10.1080/10298436.2014.943209.

[22] Y. H. Huang, *Pavement Analysis and Design*. Pearson, 2004.

[23] A. J. Alnaqbi, W. Zeiada, G. G. Al-Khateeb, K. Hamad, and S. Barakat, "Creating Rutting Prediction Models through Machine Learning Techniques Utilizing the Long-Term Pavement Performance Database," *Sustainability (Switzerland)*, vol. 15, no. 18, Sep. 2023, doi: 10.3390/su151813653.

[24] T. Tamagusko and A. Ferreira, "Machine Learning for Prediction of the International Roughness Index on Flexible Pavements: A Review, Challenges, and Future Directions," Dec. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/infrastructures8120170.

[25] Y. Li, Q. Bai, A. Hu, L. Chen, and B. Martinez-Pastor, "An adaptive updating model for pavement performance based on Deep Neural Networks," *Constr Build Mater*, vol. 449, Oct. 2024, doi: 10.1016/j.conbuildmat.2024.138391.

[26] T. Zhang, A. Smith, H. Zhai, and Y. Lu, "LSTM+MA: A Time-Series Model for Predicting Pavement IRI," *Infrastructures (Basel)*, vol. 10, no. 1, Jan. 2025, doi: 10.3390/infrastructures10010010.

[27] K. S. Basnet, J. K. Shrestha, and R. N. Shrestha, "Pavement performance model for road maintenance and repair planning: a review of predictive techniques," *Digital Transportation and Safety*, vol. 2, no. 4, pp. 253–267, 2023, doi: 10.48130/dts-2023-0021.

[28] ASTM International., *Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys (D6433-24)*. West Conshohocken, PA: ASTM International, 2024. doi: 10.1520/D6433-24.

[29] F. L. Roberts, P. S. Kandhal, E. R. Brown, D. Y. Lee, and T. W. Kennedy, *Hot Mix Asphalt Materials, Mixture Design, and Construction*, 2nd ed. NAPA Education Foundation, 1996.

[30] American Association of State Highway and Transportation Officials, *Mechanistic-Empirical pavement design guide : a manual of practice*. American Association of State Highway and Transportation Officials, 2008.

[31] J. S. Miller and W. Y. Bellinger, *Distress Identification Manual for the Long-Term Pavement Performance Program*, 5th ed. Federal Highway Administration, 2014.

[32] J. A. Deacon and C. L. Monismith, "Laboratory Flexural-Fatigue Testing of Asphalt-Concrete With Emphasis on Compound-Loading Teets," 1968.

[33] M. Ren, X. Zhang, X. Chen, B. Zhou, and Z. Feng, "YOLOv5s-M: A deep learning network model for road pavement damage detection from urban street-view imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 120, Jun. 2023, doi: 10.1016/j.jag.2023.103335.

[34] N. Shakhovska, V. Yakovyna, M. Mysak, S. A. Mitoulis, S. Argyroudis, and Y. Syerov, "Real-Time Monitoring of Road Networks for Pavement Damage Detection Based on Preprocessing and Neural Networks," *Big Data and Cognitive Computing*, vol. 8, no. 10, Oct. 2024, doi: 10.3390/bdcc8100136.

[35] H. G. R. Kerali, *Highway Development and Management: Volume 1 – Overview of HDM-4*, vol. 1. World Road Association (PIARC) and The World Bank, 2000.

[36] N. Fernando and C. Coelho, "Calibration of MEPDG Performance Models for Flexible Pavement Distresses to Local Conditions of Ontario," 2016.

[37] A. Alnaqbi, G. G. Al-Khateeb, W. Zeiada, and M. Abuzwidah, "Random forest-based frame work for multi-distress prediction in CRCP: a feature importance approach," *Discover Civil Engineering*, vol. 2, no. 1, Aug. 2025, doi: 10.1007/s44290-025-00302-z.

[38] L. Pei, T. Yu, L. Xu, W. Li, and Y. Han, "Prediction of Decay of Pavement Quality or Performance Index Based on Light Gradient Boost Machine," 2022, pp. 1173–1179. doi: 10.1007/978-3-030-81007-8_135.

[39] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," 2017.

[40] L. Lin *et al.*, "A new FCM-XGBoost system for predicting Pavement Condition Index," *Expert Syst Appl*, vol. 249, Sep. 2024, doi: 10.1016/j.eswa.2024.123696.

[41] S. Boonsiripant, C. Athan, K. Jedwanna, P. Lertworawanich, and A. Sawangsuriya, "Comparative Analysis of Deep Neural Networks and Graph Convolutional Networks for Road Surface Condition Prediction," *Sustainability (Switzerland)*, vol. 16, no. 22, Nov. 2024, doi: 10.3390/su16229805.

[42] M. M. Radwan, E. M. M. Zahran, O. Dawoud, Z. Abunada, and A. Mousa, "Comparative Analysis of Asphalt Pavement Condition Prediction Models," *Sustainability (Switzerland)*, vol. 17, no. 1, Jan. 2025, doi: 10.3390/su17010109.

[43] S. Alshawabkeh, L. Wu, D. Dong, Y. Cheng, and L. Li, “A Hybrid Approach for Pavement Crack Detection Using Mask R-CNN and Vision Transformer Model,” *Computers, Materials and Continua*, vol. 82, no. 1, pp. 561–577, 2025, doi: 10.32604/cmc.2024.057213.

[44] S. J. Pan and Q. Yang, “A survey on transfer learning,” 2010. doi: 10.1109/TKDE.2009.191.

[45] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for Transfer Learning,” 2007. [Online]. Available: http://www.cs.berkeley.edu/

[46] D. Pardoe and P. Stone, “Boosting for Regression Transfer,” 2010.

[47] S. Gupta, J. Bi, Y. Liu, and A. Wildani, “ISTRBoost: Importance Sampling Transfer Regression using Boosting,” Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.12044

[48] J. A. Prozzi and J. A. Rice, “Modeling pavement performance by combining field and experimental data,” 2001.

[49] N. Attoh-Okine and O. Adarkwa, “Pavement Condition Surveys-Overview of Current Practices,” 2013.

[50] ARPA Piemonte, “Mappa Meteoweb – rete stazione meteorologica.” Accessed: Oct. 13, 2025. [Online]. Available: https://www.arpa.piemonte.it/rischi_naturali/snippets_arpa_graphs/map_meteoweb/?rete=stazione_meteorologica#close

[51] Geoportale Piemonte, “Catalogo – Traffico giornaliero medio.” Accessed: Oct. 13, 2025. [Online]. Available: https://www.geoportale.piemonte.it/geonetwork/srv/ita/catalog.search#/search?resultType=details&sortBy=title&sortOrder=reverse&any=TRAFFICO%20GIORNALIERO%20MEDIO&fast=index&_content_type=json&from=1&to=20

[52] “Roboflow.” Accessed: Oct. 13, 2025. [Online]. Available: https://roboflow.com/

[53] J. Han, M. Kamber, and J. Pei, “Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems),” 2012.

[54] R. J. Hyndman and G. Athanasopoulos, “Forecasting principles and practice,” *University of Western Australia*, 2014.

[55] L. Breiman, “Random Forests,” 2001.

[56] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[57] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” 2017. [Online]. Available: https://github.com/Microsoft/LightGBM.

[58] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” 1967.

[59] T. O. Kvålseth, “Cautionary Note about R 2,” 1985.

[60] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, 2nd ed. Springer, 2013. [Online]. Available: http://www.springer.com/series/417