Sorbonne University

Project Report

Implementing Artificial Intelligence for Automated Data Processing in Urban Pavement Inspection for Enhanced Comfort and Safety

Elio Samaha

September 1, 2023

Contents

1 Introduction

Urban pavement systems play a critical role in ensuring safe and comfortable transportation for both pedestrians and vehicles. The condition of these pavements directly impacts factors such as ride comfort, road safety, and overall user experience. Therefore, effective inspection, assessment, and maintenance of urban pavements are essential for maintaining optimal conditions and ensuring the well-being of the public.

Traditionally, pavement inspection and assessment have been carried out through manual surveys and visual inspections. While these methods provide valuable information, they often rely on subjective judgments and can be time-consuming. Moreover, they may not fully capture subtle variations and potential issues that could impact comfort and safety.

Recent advancements in artificial intelligence (AI) and machine learning have opened up new possibilities for enhancing pavement inspection processes. By harnessing the power of AI, it becomes feasible to process and analyze large volumes of data from various sources, such as images, sensor readings, and pavement characteristics. This can lead to more accurate and efficient identification of pavement conditions, as well as predictive capabilities for future degradation.

In this project, we aim to implement an AI-driven approach for automated data processing in urban pavement inspection. By combining computer vision techniques with machine learning algorithms, we seek to develop a model capable of assessing pavement conditions, predicting potential comfort and safety issues, and providing valuable insights for effective maintenance strategies.

The subsequent sections of this report delve into our methodology and development process. In Section 2, we present a comprehensive exploration of different machine learning models for pavement comfort analysis, including K-Nearest Neighbors (KNN), Decision Trees, Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Gradient Boosting. Model comparison and similar projects' models are discussed in Section 2.2 and Section 2.3, respectively. Section 2.4 outlines the backbone of our project, providing a framework for the subsequent details. Our methodology and development approach are outlined in Section 3, discussing our model's choice, algorithm, data collection and analysis, image analysis using CNN, gradient boosting, model validation, software tool development, and hyperparameter tuning. The succeeding sections provide a comprehensive study and analysis of the application case, including data overview, collection, descriptive analysis, visualizing the data, data preprocessing, and balancing, followed by the learning phase and results. We also discuss the limitations and improvements of our approach in dealing with accuracy paradox, imbalanced datasets, and overfitting. Finally, our report concludes with reflections on the potential impact and future directions of our work in Section 5.

2 Bibliography

2.1 Different Types of Models

In this section, we will explore various machine learning and deep learning algorithms that can be employed for pavement comfort analysis in urban pavement inspection. We'll provide a brief overview of each algorithm along with relevant examples and visualizations.

2.1.1 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a simple and intuitive classification algorithm. It assigns a class label to a data point based on the majority class among its k-nearest neighbors. For the pavement comfort analysis, KNN can be used to predict comfort levels based on the features of nearby pavement sections.

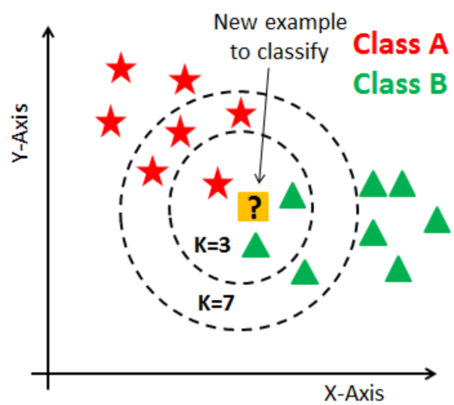

Figure 1: K-Nearest Neighbors Classification

2.1.2 Decision Trees

Decision Trees are interpretable and can capture complex relationships between features and outcomes. Each internal node represents a feature, and each leaf node represents a class label. In this project, we can design a decision tree to predict pavement comfort levels based on various features, such as microtexture and environmental conditions.

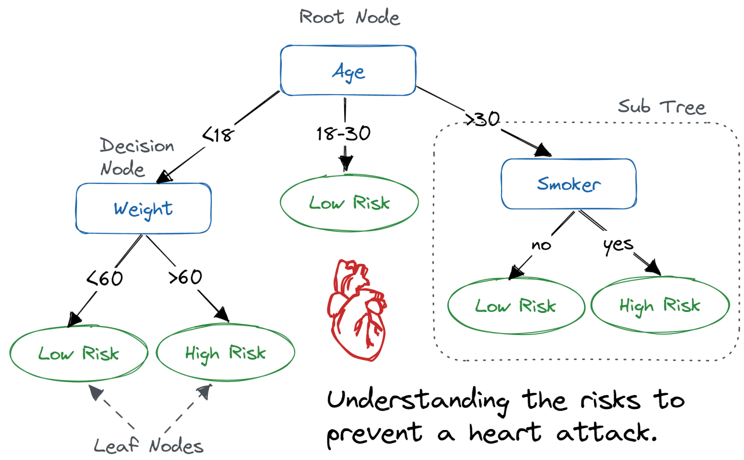

Figure 2: Example Decision Tree for Comfort Prediction

2.1.3 Deep Neural Networks (DNN)

Deep Neural Networks are powerful models capable of learning complex patterns. They consist of multiple hidden layers that transform input features into predictions. In this project, we can use DNNs to capture intricate relationships between pavement characteristics and comfort levels.

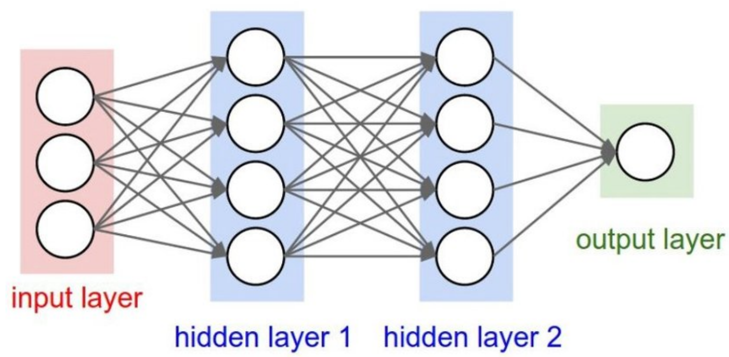

Figure 3: Deep Neural Network Architecture

2.1.4 Convolutional Neural Networks (CNN)

Convolutional Neural Networks excel in image analysis tasks. They use convolutional layers to extract spatial features from images. In your project, CNNs can be applied to analyze pavement images and extract relevant features for predicting comfort levels.

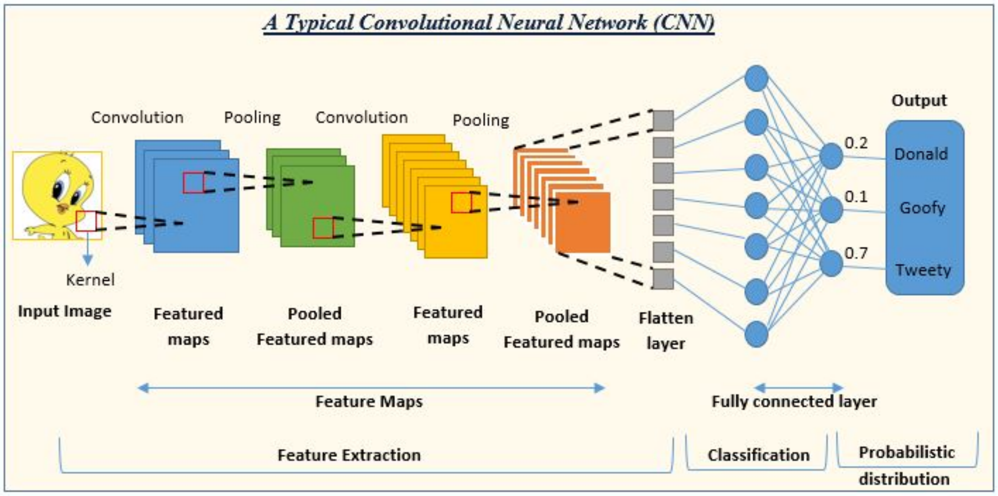

Figure 4: Convolutional Neural Network Architecture

2.1.5 Gradient Boosting

Gradient Boosting is an ensemble learning method that combines weak learners to create a strong predictive model. It builds trees sequentially, where each tree corrects the errors of the previous one.Gradient Boosting can be used to predict pavement comfort levels by combining information from different pavement features.

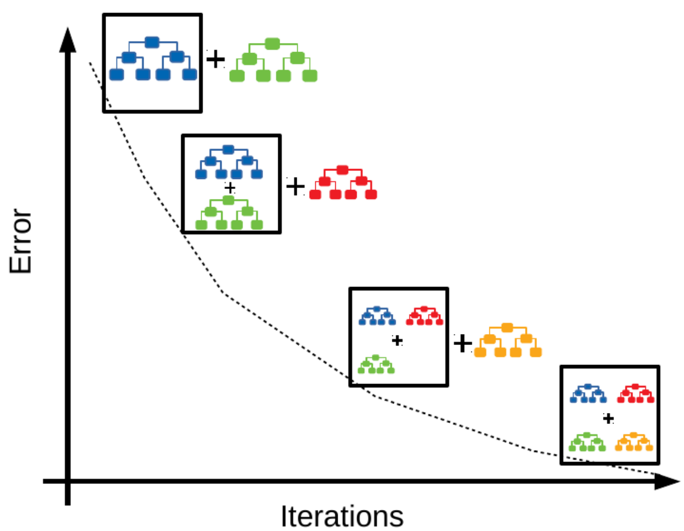

Figure 5: Gradient Boosting Ensemble [7].

Incorporating these diverse machine learning and deep learning algorithms can provide a comprehensive approach to predicting pavement comfort levels, considering both numerical features and image data.

2.2 Model Comparison

In this section, we will compare different machine learning and deep learning models that can be utilized for pavement comfort analysis in urban pavement inspection. Each model has its advantages and disadvantages, and we will evaluate their suitability for your project.

Table 1: Comparison of Different Machine Learning Models for Pavement Comfort Analysis

Model	**Advantages**	**Disadvantages**
K-Nearest Neighbors (KNN)	Simple to understand and implement, No assumptions about data, Can be good if pavement features have close neighbors with similar comfort levels	Computationally expensive, Sensitive to irrelevant features and the scale of the data, May not work well with a large number of features (e.g., various pavement characteristics, environmental images, etc.)
Decision Trees	Simple to understand, Can handle both numerical and categorical data, May work well with a mix of pavement features	Prone to overfitting, Can be unstable with small variations in data, May not capture complex relationships between pavement characteristics and comfort level
Deep Neural Networks (DNN)	Can model complex patterns, Good with large amount of data, Could be useful for capturing complex relationships between pavement characteristics and comfort level	Require a lot of data, Computationally expensive, Hard to interpret, May overfit with less data
Convolutional Neural Networks (CNN)	Good for image data, Can capture spatial features, Could be useful if environmental images are important features	Require a lot of data, Computationally expensive, Hard to interpret, May not work well if images are not important features
Gradient Boosting	Can capture complex patterns, Good with structured data, Often provides excellent accuracy, Could work well with a mix of pavement features	Can be computationally intensive, May require careful tuning of parameters, Overfitting could be a concern

Continued on next page

Table 1 – continued from previous page

Model	Advantages	Disadvantages
Deep Learning (General)	Can model complex patterns, Good with large amount of data, Could capture complex relationships between features and comfort level	Requires large amounts of data, Computationally intensive, Model interpretability can be difficult, Overfitting could be a concern

In this section, we compared different machine learning and deep learning models for their suitability in pavement comfort analysis for urban pavement inspection. Each model has its own strengths and limitations, and the choice of model depends on the specific project requirements. After evaluating various models based on their advantages and disadvantages, we have decided to adopt a hybrid approach that combines Convolutional Neural Networks (CNN) and Gradient Boosting.

2.3 Similar projects' models

2.3.1 Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks

The study *Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks* by Sudakov, O., Burnaev, & Koroteev (2019) explores the application of machine learning techniques for predicting permeability utilizing 3D scans of Berea sandstone subsamples [1].

The study presents a study on the use of deep learning techniques for the prediction of permeability in sandstone samples imaged with X-ray microtomography. The study focuses on the use of Minkowski functionals and Deep Learning-based descriptors of 3D images and 2D slices as input features for predictive model training and prediction. The study also evaluates the use of Gradient Boosting (XgBoost) and various architectures of Deep Neural Networks (DNNs) as regression methods to assess the predictive power of the generated features.

The study generated a training set containing 3D images of sandstone samples imaged with X-ray microtomography and corresponding permeability values simulated with Pore Network approach. The results show that the combination of CNN and XgBoost provides the best performance for the prediction of permeability, with a mean absolute error (MAE) of 0.0007 and a coefficient of determination (R^2) of 0.98. The generated features are effective in capturing the microstructure information of the materials, and the study demonstrates the potential of deep learning techniques for the prediction of permeability in sandstone samples.

The study also compares the predictive power of various feature sets and methods, and shows that the use of generated features with DNNs and XgBoost provides better performance than the use of raw images or hand-crafted features. The study demonstrates the applicability of machine learning for image-based permeability prediction and opens a new area of Digital Rock research.

Overall, the study provides insights into the use of deep learning techniques for the prediction of permeability in sandstone samples, and demonstrates the potential of generated features with DNNs and XgBoost for this task. The study also highlights the importance of feature engineering and model selection in machine learning applications, and provides a framework for future research in this area.

2.3.2 A New Hybrid Convolutional Neural Network and eXtreme Gradient Boosting Classifier for Recognizing Handwritten Ethiopian Characters

A similar project titled *A New Hybrid Convolutional Neural Network and eXtreme Gradient Boosting Classifier for Recognizing Handwritten Ethiopian Characters* by Halefom Tekle Weldegebriel[a], Han Liu[b] (Member IEEE), Anwar Ul Haq[a], Emmanuel Bugingo[a], and Defu Zhang[a*] (Member IEEE) was conducted.

[a]School of Informatics, Xiamen University, Xiamen 361005, China,
[b]School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.
*Corresponding author: Defu Zhang (dfzhang@xmu.edu.cn).

In this study, the authors combined a Convolutional Neural Network (CNN) and eXtreme Gradient Boosting (XGBoost) to create a hybrid model for recognizing handwritten Ethiopian characters. The CNN was used as a trainable automatic feature extractor from the raw images, and XGBoost took the extracted features as input for recognition and classification. The architecture of the hybrid CNN-XGBoost model was designed by swapping the last fully connected layer of the CNN model with an XGBoost classifier. The input of the activation function was the linear combination of the outputs from the previous hidden layer with trainable weights, plus a bias term.

The input to the model was a dataset of handwritten Ethiopian characters. The output of the model was the classification of the characters. The authors used the XGBoost algorithm, which is an integrated learning algorithm based on gradient boosting, to efficiently achieve higher classification accuracy.

The results of the experiments showed that the hybrid CNN-XGBoost model achieved an error rate of 0.1612, while the CNN with a fully connected layer achieved an error rate of 0.4630. This suggests that XGBoost as a classifier performs better than the traditional fully connected layer[2].

2.3.3 Gradient Boosting Machine and Object-Based CNN for Land Cover Classification

Bui et al. [3] presented a novel method for land cover classification using a combination of Convolutional Neural Networks (CNNs) and various gradient boosting algorithms, specifically XGBoost, LightGBM, and Catboost. The study used SPOT7 imagery as the input data. The proposed algorithm involved several stages, including image segmentation, feature extraction, normalization, and graph generation, culminating in the training of the hybrid model. The performance of these CNN-based gradient boosting methods surpassed other benchmarked methods, achieving an overall accuracy ranging from 0.8905 to 0.8956. This work demonstrates the potential of employing object-based image analysis and CNN-based gradient boosting algorithms in enhancing the accuracy of land cover analysis.

2.3.4 Impact of Factors on Adhesive Behavior of Asphalt: A Hybrid BOA-SVR Model Approach

In the study titled "Impact of Factors on Adhesive Behavior of Asphalt: A Hybrid BOA-SVR Model Approach" by M. Arifuzzaman, M. Aniq Gul, K. Khan, and S. M. Zakir Hossain, the researchers investigate the adhesive behavior of asphalt and develop a hybrid model using the Bayesian Optimization Algorithm (BOA) and Support Vector Regression (SVR) techniques. The aim of the study is to predict the adhesive force of asphalt considering various factors such as conditions, binder types, and Carbon Nanotube (CNT) doses.

Experimental data comprising 405 observations of asphalt adhesive force under different factor combinations were collected. The researchers utilized Atomic Force Microscopy (AFM) to estimate the adhesive force. The collected data were used to develop the hybrid BOA-SVR model.

The results demonstrate that the developed hybrid BOA-SVR model accurately predicts the adhesive force of asphalt. The model shows a strong fit between the experimental and predicted values, with a high coefficient of determination (R2) and adjusted R2 values. Additionally, the statistical error parameters, including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), indicate the model's accuracy in predicting adhesive behavior.

The study concludes that the hybrid BOA-SVR model provides a reliable and robust approach for estimating the adhesive force of asphalt under different conditions, binder types, and CNT doses. The findings contribute to a better understanding of asphalt properties and have implications for improving asphalt pavement design and performance.[4]

2.3.5 Automatic Pothole Detection in Asphalt Pavement Using Steerable Filter and Artificial Intelligence

In the study titled *Automatic Pothole Detection in Asphalt Pavement Using Steerable Filter and Artificial Intelligence*, the authors propose a comprehensive approach for automated pothole detection in asphalt pavement. The approach, named Steerable Filter and Artificial Intelligence-Based Pothole Detection Model (SF-AI-PDM), consists of three main modules: image acquisition and feature extraction, data set construction, and AI model training and prediction.

The image processing stage utilizes Gaussian Filter (GF) for denoising the original digital image and Steerable Filter (SF) for generating a pothole-resilient map. Integral Projection (IP) analysis is then applied to extract features characterizing potholes. Additionally, a moving average technique is employed to reduce the dimensionality of the extracted features.

The SF-AI-PDM incorporates two AI methods, namely Artificial Neural Network (ANN) and

Least Squares Support Vector Machine (LS-SVM), for training and predicting the pothole classification model. Evaluation metrics such as classification accuracy rate (CAR), true positive rate (TPR), false positive rate (FPR), false negative rate (FNR), and true negative rate (TNR) are used to assess the performance of the models. Experimental results indicate that LS-SVM outperforms ANN, achieving higher CAR and AUC values.

The proposed SF-AI-PDM shows promise for effective pothole detection in asphalt pavement, offering potential benefits for transportation agencies and road inspectors. Future research directions include exploring advanced AI methods, ensemble learning strategies, and integrating sophisticated image processing techniques to enhance feature extraction and pothole size estimation.[5].

2.3.6 Utilization of Computer Vision Strategies in Pavement Distress Engineering: A Review

In this review, the authors explored the utilization of computer vision strategies in pavement distress engineering. The study focused on the application of computer vision techniques, such as digital image processing and computer vision algorithms, for analyzing asphalt structures, characterizing aggregates, and monitoring macroscopic-scale pavement conditions. The importance of computer vision in improving detection efficiency and accuracy of pavement distress, particularly cracks, was discussed.

The review highlighted the use of various techniques for crack detection and segmentation, including thresholding techniques, deep learning algorithms, and edge detection methods. The authors discussed the limitations of computer vision techniques, such as the presence of external noise and non-uniform illumination in acquired images, and the strategies employed to address these limitations, such as filtering and mean and standard deviation values.

Furthermore, the review discussed the utilization of ground-penetrating radar (GPR) for pavement distress detection, as well as classification techniques for crack types. The use of artificial intelligence (AI) techniques, such as support vector machines (SVM), artificial neural networks (ANN), and deep learning algorithms, for crack classification was explored.

The review also highlighted the detection of potholes using computer vision techniques, including the integration of drones with image processing for efficient and accurate detection. The authors emphasized the potential of AI-enhanced drones in healthcare, specifically for transporting medications, treatments, and biological samples to inaccessible areas.

Overall, the review provided insights into the current state-of-the-art work of computer vision strategies in pavement distress engineering, highlighting the advancements, challenges, and future directions in this field[6].

2.3.7 Summary Table

Table 2: Summary of Similar Projects

Project	**Field of Application**	**Model Used**	**Results**	**Adapted to Our Project**
Driving digital rock towards machine learning	Permeability prediction using 3D scans of sandstone	Gradient boosting, Deep neural networks	Achieved MAE of 0.0007, R^2 of 0.98 for permeability prediction	Potential for capturing microstructure information
A New Hybrid Convolutional Neural Network and eXtreme Gradient Boosting Classifier	Handwritten Ethiopian character recognition	CNN, XGBoost	Achieved error rate of 0.1612	Yes, superior performance over CNN-only
Gradient Boosting Machine and Object-Based CNN	Land cover classification using imagery	CNNs, XGBoost, LightGBM, Catboost	Achieved accuracy ranging from 0.8905 to 0.8956	Enhanced accuracy in land cover analysis
Impact of Factors on Adhesive Behavior of Asphalt	Adhesive behavior prediction of asphalt	Bayesian Optimization Algorithm (BOA), SVR	Accurate prediction of adhesive force	Yes, accurate prediction under varying factors
Automatic Pothole Detection in Asphalt Pavement	Pothole detection using image processing	Steerable Filter, Artificial Intelligence	Comprehensive approach for automated pothole detection	Yes, promising approach for pothole detection
Utilization of Computer Vision Strategies in Pavement Distress Engineering	Pavement distress analysis using computer vision	Various computer vision techniques	Enhanced detection of pavement distress	Insightful strategies for crack detection

2.4 Project's Backbone

In this section, we will discuss a similar project to ours that served as inspiration for adding AI components to our model. The project is called *Alertinfra*, a software designed for the detection of potentially hazardous situations. It was developed in collaboration between DES (Division Exploitation et Sécurité) and LRPC (Laboratoire Régional des Ponts et Chaussées) as part of a research operation titled *Sécurité des Itinéraires* (Route Safety). *Alertinfra* serves as an aid in diagnosing the safety of inter-urban routes.

The software operates by analyzing data collected from RST (Réseau Scientifique et Technique) measurement devices. It identifies specific zones where alerts are triggered based on the analysis of this data.

The collected information includes color video images to characterize the road environment, physical measurements related to the road's geometry (turning radius, slope, camber), and measurements related to the surface characteristics of the road (adhesion, macrotexture, vertical acceleration).

The main goal of the *ALERTINFRA* software is to analyze the results of infrastructure parameters' measurements, primarily focusing on geometry (turning radius, camber, slope), and surface characteristics (adhesion, macrotexture, uniformity). It specifically applies to inter-urban areas.

The measured parameters are recorded at intervals of 1 meter. Radii are given in meters (positive for left turns, negative for right turns), slopes and camber in percentage, adhesion in SCRIM CFT equivalent (coefficient of transversal friction), macrotexture in HS equivalent (height to sand), and a uniformity indicator is given in "g" (acceleration of gravity), measuring the acceleration of unsuspended masses of the vehicle.

The Alert Index: Based on the statistical weights defined earlier, it is possible to identify an alert index for each turn. The combination of different weights of triggered alerts for a given turn allows the calculation of the alert index (IA). This index helps classify turns in descending order of dangerousness.

$$IA = \frac{\sum(weight \times alert)}{\sum weight} \text{ or } \frac{\sum(weight \times alert)}{0.217}$$

For an entire itinerary, this index serves to prioritize interventions for the manager. However, it is important to note that no weight has been defined for two alerts in a turn. The alert index, as currently calculated, is therefore partial (for alerts V1 to V13) and will need to be supplemented based on future work.

The output of the product is a result file in .csv format, directly openable in Excel. It contains:

- the division of the measurement section into straight-line sections (with radii of curvature at 500, 600, or 700 m) and turns,
- indicating the start and end PRs of the sections, the radius of curvature of each section, and its length,
- a column list of triggered alerts with more precise positioning of straight-line alerts and an indication of the different alert levels for the relevant alerts,
- the calculation of the resulting coefficient from the weighting of different alerts.

While *Alertinfra* provides valuable insights by analyzing data collected from various measurements, it also comes with certain limitations that underscore the potential for improvement through the integration of Artificial Intelligence (AI).

Alertinfra's approach relies heavily on predefined thresholds and weighted alerts to identify potentially dangerous sections of road. However, this approach may have limitations in capturing complex patterns and nuanced relationships between various road parameters and safety risks. Additionally, the current methodology might struggle with scenarios where the impact of combined factors is not well-represented by predefined thresholds.

In contrast, the incorporation of AI into our project brings the advantage of more sophisticated pattern recognition and predictive capabilities. AI models, especially those leveraging deep learning techniques, have the potential to learn intricate patterns from vast datasets, allowing them to adapt to a wider range of road conditions and safety concerns. This is particularly relevant in urban pavement inspection, where the interplay of multiple variables can significantly impact pavement comfort and safety.

By utilizing AI, our project aims to enhance the accuracy and granularity of safety assessments by considering a broader spectrum of factors and their interactions. This improvement holds the promise of not only identifying hazardous areas more effectively but also offering a more comprehensive understanding of the underlying factors contributing to safety concerns. Ultimately, the integration of AI into pavement inspection represents a significant leap forward in harnessing technology to ensure safer and more comfortable urban road networks.

3 methodology and development

3.1 Our Model's Choice

Based on our literature review and the comparative analysis of various machine learning models, we decided to opt for a hybrid approach that combines the strengths of Convolutional Neural Networks (CNN) and Gradient Boosting methods. This decision is motivated by several factors:

3.1.1 Efficient Image Feature Extraction using CNN

The CNN model has shown its strength in dealing with image data, as we are dealing with pavement surface images in our project. It is particularly good at capturing spatial features from images, which are crucial for pavement comfort analysis. This is confirmed by the similar projects we reviewed, notably the project by Bui et al. [3] that used CNN-based gradient boosting methods for land cover classification, and the project by Weldegebriel et al. [2] which used a hybrid CNN-XGBoost model for recognizing handwritten Ethiopian characters.

3.1.2 Robust Classification using Gradient Boosting

On the other hand, the Gradient Boosting model is excellent with structured data and is known to provide superior accuracy. It can capture complex patterns and is suitable for working with a mix of pavement features we aim to use for predicting the comfort levels. This technique's efficiency is supported by Sudakov et al.'s study [1], which found a combination of CNN and Gradient Boosting (XgBoost) to be the best performer for permeability prediction in sandstone samples.

3.1.3 Hybrid Model's Advantages

The hybrid model that uses CNN for feature extraction and Gradient Boosting for classification combines the strengths of these two approaches. CNN efficiently processes our image data, extracting meaningful features, which are then combined with the other pavement features for the final prediction using Gradient Boosting. This model provides a good balance between computational complexity and predictive power, and it allows us to exploit both the spatial relationships present in the image data and the structured relationships in the measured pavement features.

In conclusion, this hybrid approach appears to be the most suitable for our project, as it combines the benefits of efficient image feature extraction using CNN and robust classification using Gradient Boosting methods.

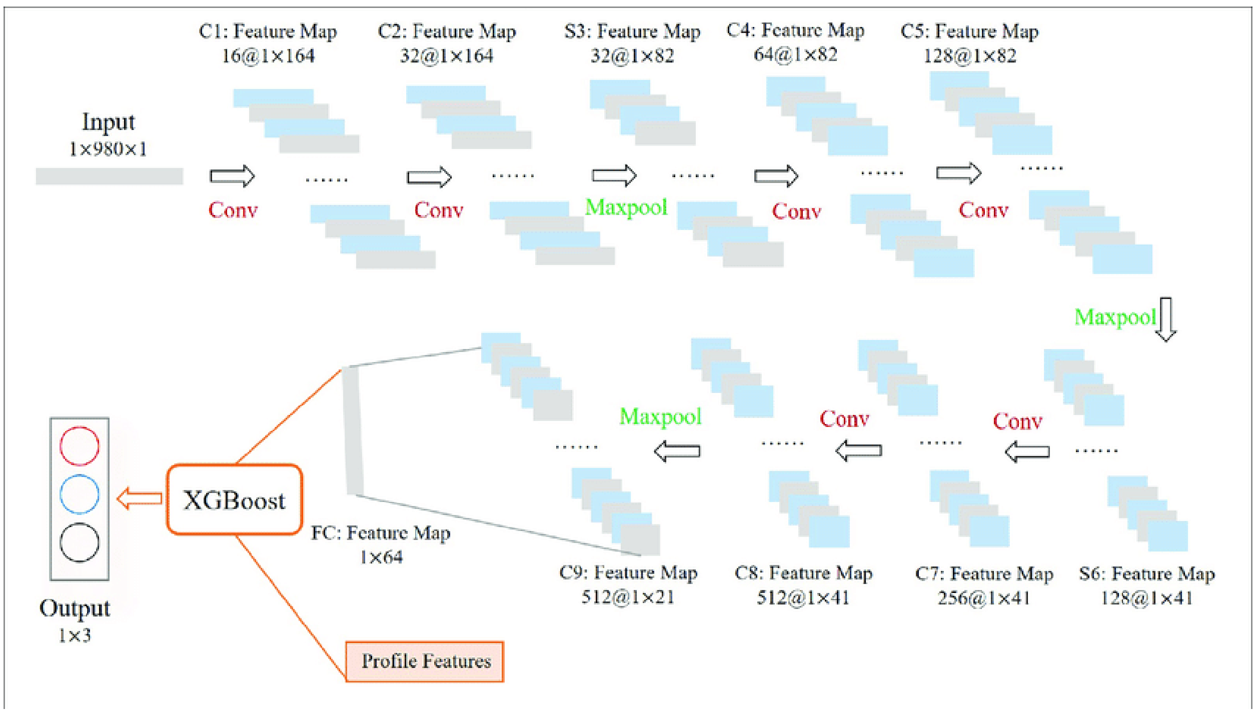

Figure 6: CNN and gradient boosting combined [8]

3.2 Our Model's Algorithm

The development and implementation of the machine learning model involves a number of steps, which are described below:

3.2.1 Data Collection and Analysis

The tasks begin with the collection of data which could be from devices that measure various parameters of the pavement, including possibly the texture, roughness, temperature, etc., and images of the pavement surfaces.

3.2.2 CNN for Image Analysis

Convolutional Neural Networks (CNN) are used to analyze the images of the pavement surfaces. These networks can learn to recognize complex patterns in the image data, which can be valuable for assessing the condition of the pavement. The output from the hidden layers of the CNN (often the fully-connected layers towards the end of the network) is used as a compact representation of these patterns.

3.2.3 Gradient Boosting for Final Prediction

The numerical data from the measurements, as well as the image-based features extracted by the CNN, are used as input to a Gradient Boosting model. Gradient Boosting models are effective for a range of classification and regression tasks, and can handle a mix of categorical and numerical data. In this case, the target variable is the "comfort level" of the pavement.

3.2.4 Model Validation

Cross-validation or a separate test set is used to evaluate the performance of the model. It's important to optimize the model to improve its accuracy and robustness, which might involve tuning the parameters of the CNN, the Gradient Boosting model, or the feature extraction process.

3.2.5 Software Tool Development

The entire process is encapsulated in a software tool. This tool automates the data analysis and prediction process, and potentially provides a user-friendly interface for non-experts to use the model. It should be able to handle the specifics of the in-situ measurements and the calculation of the comfort levels of urban pavement coatings.

3.2.6 Tuning hyperparameters

Both CNNs and Gradient Boosting models have several hyperparameters that need to be tuned, such as the architecture of the CNN (e.g., number and size of the convolutional layers) and the parameters of the Gradient Boosting model (e.g., learning rate, tree depth). This tuning process is done through cross-validation.

Pre-processing of the data before inputting it into the models is crucial. For the CNN, this might involve resizing the images, normalizing the pixel values, etc. For the Gradient Boosting model, it might involve scaling numerical variables, encoding categorical variables, handling missing values, etc.

Finally, it's important to ensure that there is enough data to train the models, especially the CNN which can require large amounts of data to learn effectively. If data is scarce, techniques like data augmentation for the CNN and bootstrap aggregation for the Gradient Boosting model could be considered.

4 Study and Analysis of Application case

4.1 Data

Our project's foundation lies in a comprehensive collection of road measurements accompanied by corresponding images. This amalgamation of data serves as the bedrock for our subsequent analysis and modeling endeavors.

4.1.1 Dataset Overview

The dataset employed in this project is a product of meticulous measurements conducted along a specific route – the Montaigu to Lons stretch in Department 39. On the 13th of August 2012, measurements were executed across a section spanning from PR 21 at the outset to PR 12 at the culmination and conversely. This examination was limited to lane G of the D678 road, and measurements were acquired at regular 1-meter intervals along the road's expanse of approximately 8924 meters. The dataset encompasses the entire extent of this road section, furnishing an exhaustive source of information for subsequent analysis and modeling endeavors.

4.1.2 Data Collection

Our dataset comprises information from 17,842 sampled points along the road, collected at intervals of 1 meter. Additionally, a collection of 3569 images was captured at 5-meter intervals. These images provide visual insights into road conditions, while measurements encapsulate a range of crucial parameters: radius of curvature, adhesion, vertical acceleration, velocity (excluding car speed), road slope, road cant, and road roughness. This multidimensional dataset forms the cornerstone of our predictive model, combining numerical metrics with visual cues.

4.1.3 Descriptive Analysis

To develop a comprehensive understanding of the dataset's characteristics, we conducted a preliminary descriptive analysis. This exploration facilitated a high-level comprehension of the dataset's attributes, enabling the identification of patterns and trends. Our analysis unveiled substantial variability in each parameter, underscoring the diverse road conditions that our model is designed to encompass.

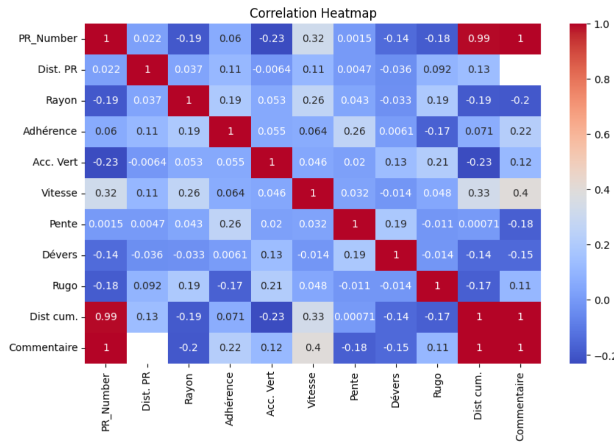

Figure 7: Correlation heat map

	Rayon	Adhérence	Acc. Vert	Vitesse	Pente	Dévers	Rugo
count	17842.000000	17842.000000	17842.000000	17842.000000	17842.000000	17842.000000	17842.000000
mean	437.102567	0.667001	0.479072	39.264432	0.023215	1.435540	0.889760
std	289.414965	0.073484	0.178581	4.146769	3.261888	2.392006	0.290422
min	-598.000000	0.458000	0.100000	16.000000	-5.800000	-7.600000	0.330000
25%	302.000000	0.614000	0.370000	37.000000	-2.200000	0.200000	0.690000
50%	600.000000	0.667000	0.420000	40.000000	0.000000	1.700000	0.830000
75%	600.000000	0.722000	0.530000	42.000000	2.300000	2.800000	1.030000
max	600.000000	0.884000	1.680000	46.000000	5.600000	8.000000	2.180000

4.1.4 Visualizing the Data

Figure 8 showcases a sample image from the dataset, providing a glimpse into the visual information complementing our quantitative measurements. Furthermore, Figure 9 illustrates the distribution of key parameters, depicting their variability and potential insights.

Figure 8: road image with shape (1040, 1392, 3)

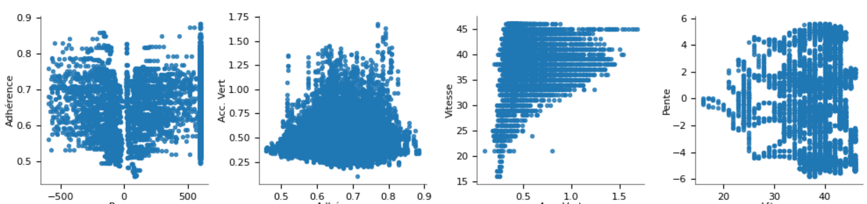

Figure 9: Distribution of Key Parameters

N° PR	Dist. PR	Rayon	Adhérence	Acc. Vert	Vitesse	Pente	Dévers	Rugo	Dist cum.	Evénement	Commentaire	Situation
N° PR	Dist. PR	Rayon	Adhérence	Acc. Vert	Vitesse	Pente	Dévers	Rugo	Dist cum.	Evénement	Commentaire	Situation
12	0	22	730	20	23	2.1	2.5	96	1	PR	12.0	En agglo
12	1	23	730	20	23	2.5	2.2	65	2	nan	nan	En agglo
12	2	23	730	30	23	2.8	2.5	71	3	nan	nan	En agglo
12	3	23	730	30	23	2.9	2.4	75	4	nan	nan	En agglo
12	4	23	730	24	23	2.6	2.2	75	5	nan	nan	En agglo

Figure 10: DataFrame Head

4.1.5 Data Preprocessing and Balancing

In addition to the tabular data, our dataset includes images captured at regular intervals along the road section. Prior to their integration into the analysis pipeline, the images underwent preprocessing to ensure compatibility with our models. This preprocessing involved resizing the images to a standardized resolution of 224x224 pixels and normalizing pixel values to a range of [0, 1]. This step enhances the images' interpretability by machine learning algorithms and helps reduce the computational burden during training. Furthermore, due to the difference in the number of available images and road samples, we adopted a strategic approach to ensure balanced representation for both data types. To address the disparity in the number of road samples and images, a strategic transformation was employed. The original dataset, composed of 17,842 samples each with 7 features, was transformed to ensure compatibility. By aggregating every five consecutive samples, the dataset was distilled to a more balanced representation, ultimately yielding a 3,569 x 7 DataFrame. This transformation entailed calculating the mean value for each feature within these groups of samples. Consequently, the refined dataset achieves equilibrium between samples and images, a crucial factor for optimal model performance. This approach retains essential information while harmonizing the overall composition, paving the way for a more effective predictive modeling process.

Rayon	Adhérence	Acc. Vert	Vitesse	Pente	Dévers	Rugo
Rayon	Adhérence	Acc. Vert	Vitesse	Pente	Dévers	Rugo
20.0	0.713	0.4	21.0	2.1	2.1	0.37
20.0	0.713	0.8	21.0	1.9	2.1	0.53
20.0	0.713	0.1	21.0	1.3	1.6	0.34
20.0	0.713	0.2	21.0	0.8	2.2	0.59
21.0	0.713	0.36	21.0	1.3	2.0	0.45

Figure 11: preprocessed Dataframe

Figure 12: preprocessed image with shape (224, 224, 3)

This comprehensive dataset, encompassing both quantitative measurements and visual data, empowers our project to build a robust predictive model for road safety assessment.

4.2 Learning Phase

The learning phase of our project encountered a significant challenge posed by the inherent imbalance in the dataset—wherein there were more samples than images. To tackle this issue and effectively utilize the strengths of both data types, we adopted an innovative approach that capitalized on their respective attributes, while also addressing potential biases.

Given the limited number of available images, our strategy involved a novel grouping technique. We aggregated five consecutive samples to correspond with each image, ensuring a harmonized representation of image and sample data. This unique aggregation approach allowed us to harness the power of visual information alongside measurements, facilitating a more comprehensive understanding of the data.

In the pursuit of constructing a powerful predictive model, we ventured to create a hybrid solution by integrating Convolutional Neural Networks (CNNs) and Gradient Boosting algorithms. The CNN component played a pivotal role in feature extraction from images, producing a condensed representation of their visual attributes. These extracted features were then concatenated with the vector containing measurements from the associated samples. The amalgamation of these diverse features formed the input for the Gradient Boosting algorithm. This integration fostered intricate learning between visual characteristics and measurements, resulting in a model capable of nuanced insights and robust predictions.

Our pioneering approach leverages both image and sample data to not only enhance predictive accuracy but also to enrich the understanding of complex real-world scenarios. This unique fusion underscores our commitment to solving real-world challenges by effectively exploiting the diverse range of data sources and leveraging advanced techniques.

4.3 Results

The results of our approach reveal promising performance, indicating the efficacy of our novel methodology. The integration of image and sample data has led to a comprehensive model that demonstrates strong predictive capabilities. The following are the outcomes of our experimentation:

- **Validation Fold Metrics for XGBoost:**
  - Accuracy: 99.56%
  - Precision: 99.10%
  - Recall: 100%
  - F1 Score: 99.55%
- **Average Precision Score:** 99.74%
- **ROC-AUC Score:** 99.70%

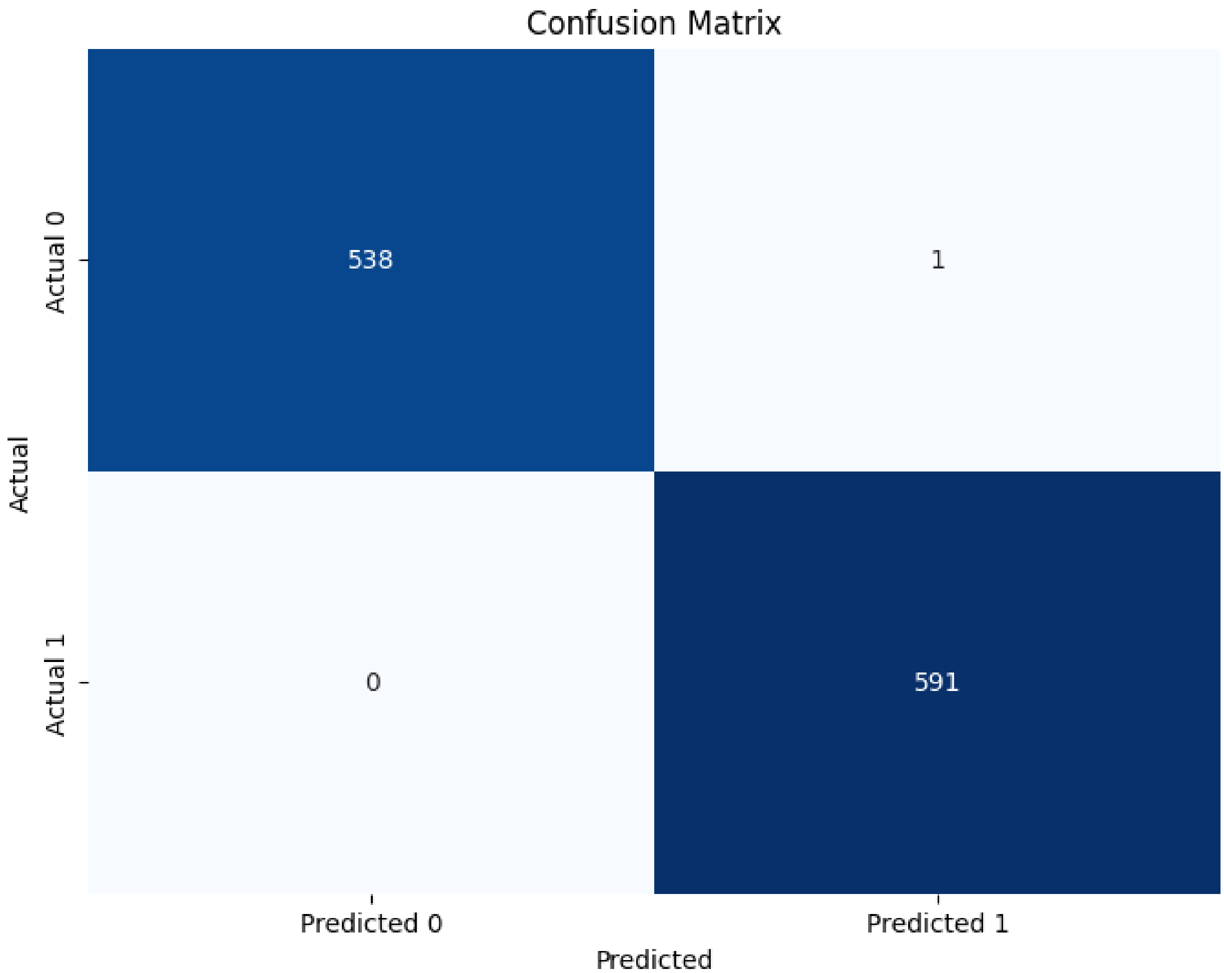

Figure 13: confusion matrix for validation set

These metrics reflect the model's ability to make accurate predictions and capture the underlying patterns within the data. However, it's crucial to address potential concerns of overfitting, especially considering the high validation performance.

We acknowledge that the model's performance on the test set did not yield similar favorable results:

- **Test Set Metrics:**
  - Accuracy: 98.88%
  - Precision: 0%
  - Recall: 0%
  - F1 Score: 0%

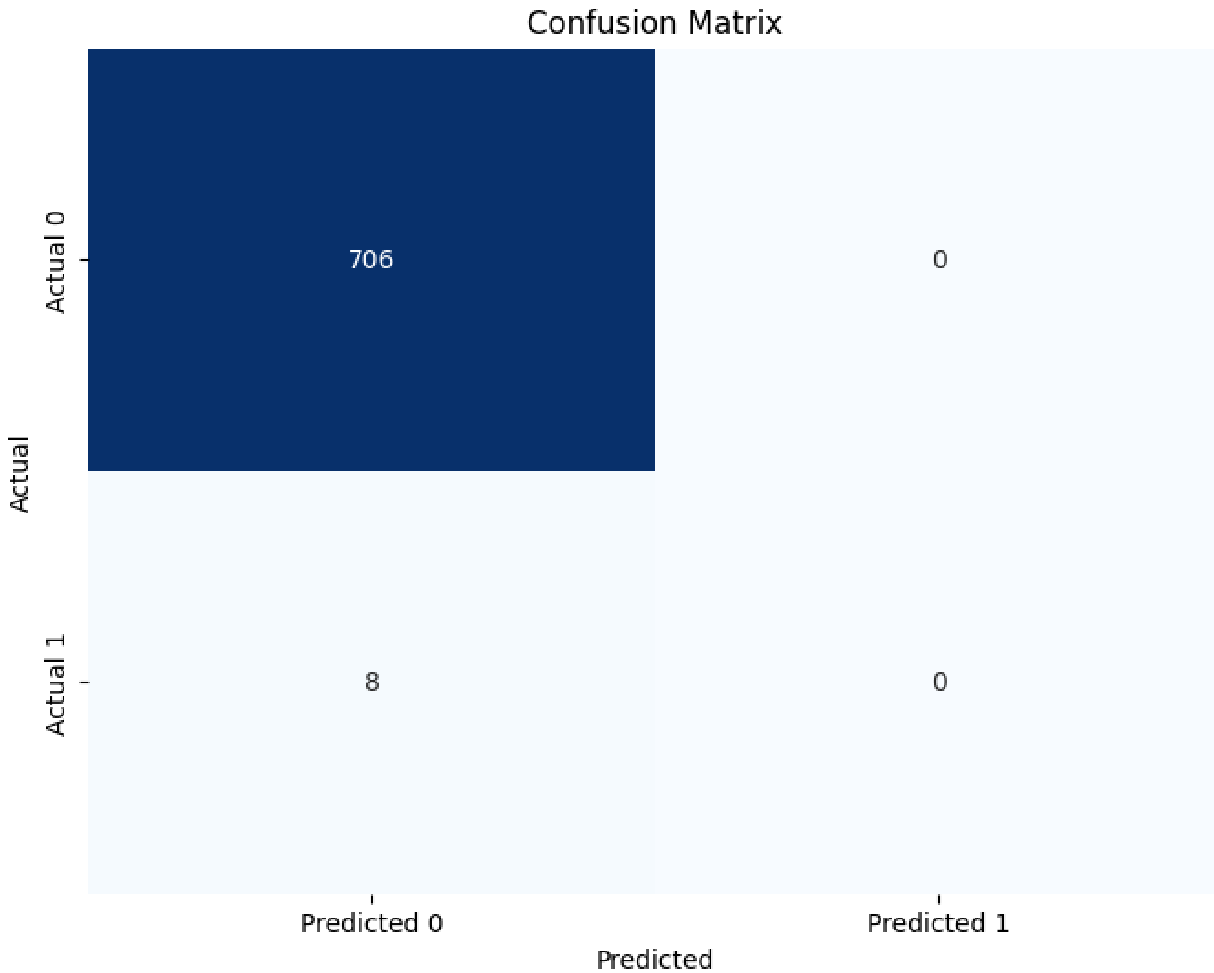

Figure 14: confusion matrix for test set

This disparity between validation and test results necessitates a careful examination of potential overfitting. It's important to investigate the sources of discrepancies and consider employing strategies to mitigate overfitting, such as regularization techniques, feature engineering, and model complexity optimization.

In conclusion, our pioneering approach, which combines the strengths of image and sample data, holds immense promise in addressing complex challenges. While the results are encouraging, further efforts are needed to ensure the model's robustness and generalization capabilities on unseen data. Our commitment remains unwavering as we continue to explore innovative solutions at the intersection of diverse data sources and advanced algorithms.

4.4 Limits and Improvements

While our approach shows promising results, there are certain limitations and areas for improvement that need to be acknowledged. These aspects highlight the complexity of dealing with class imbalance and the need for continued refinement:

4.4.1 Accuracy Paradox and Imbalanced Datasets

It's evident that the high accuracy achieved in cross-validation doesn't translate into strong performance on the test set, particularly for the positive class (class 1). This phenomenon is known as the "accuracy paradox," commonly observed in scenarios with imbalanced datasets. The majority class's prevalence can lead to misleadingly high accuracy while the minority class's performance suffers. Here are strategies to address this challenge:

- **Use Different Evaluation Metrics:** Relying solely on accuracy is inappropriate for imbalanced datasets. Metrics like precision, recall, and F1-score provide a clearer picture of model performance. Metrics such as average precision score and ROC-AUC score offer a more comprehensive evaluation.
- **Threshold Adjustment:** Consider adjusting the classification threshold to achieve a better balance between precision and recall. This modification can be based on the business requirements and the model's intended application. Utilize the `predict_proba` method to fine-tune the threshold.
- **Class Weighting:** XGBoost's `scale_pos_weight` parameter allows you to assign greater weight to the minority class, aiding the model in effectively learning from the underrepresented class.
- **Ensemble Methods:** Techniques like Random Forest and AdaBoost, which emphasize the minority class during training, can be advantageous in enhancing model performance.
- **Feature Engineering and Selection:** Optimize the input features by eliminating irrelevant or redundant attributes. This can reduce noise in the data and subsequently improve model performance.

4.4.2 Overfitting Considerations

The impressive metrics achieved, particularly the Average Precision Score and ROC-AUC Score, also warrant careful scrutiny. While they reflect model strength, they can also signal overfitting, as evidenced by the disparity between validation and test results. Here's how to address potential overfitting:

- **Evaluate on Unseen Data:** The importance of testing the model on an entirely new and unseen test set cannot be overstated. A significant drop in performance from validation to test is a red flag for overfitting.
- **Regularization:** Techniques like dropout, L2 regularization, and early stopping can counter overfitting by introducing constraints on the model's complexity.
- **Model Simplification:** Complex models are more prone to overfitting. Consider simplifying the model structure to avoid capturing noise in the training data.
- **Feature Engineering:** Ensure the features selected are informative and not contributing to noise. Discarding irrelevant features can help alleviate overfitting.

- **Cross-Validation:** Implement cross-validation with multiple folds to obtain a more robust assessment of the model's performance. Consistency across different folds signifies more reliable results.
- **Hyperparameter Tuning:** Thoroughly review hyperparameters to confirm they are well-calibrated and not promoting noise fitting.
- **Data Augmentation or Collection:** If feasible, consider expanding the dataset by augmenting existing data or collecting additional samples. More data can mitigate overfitting by enabling the model to learn more diverse patterns.

In conclusion, our model's performance, though promising, is accompanied by the complexities of addressing class imbalance and overfitting. The iterative nature of model development underscores the need for continuous refinement and adaptation. By incorporating these considerations and making informed decisions, we aim to create a model that strikes a balance between high performance on validation data and robust generalization to new and unseen data.

5 Conclusion

The significance of urban pavement systems in ensuring safe and efficient transportation cannot be overstated. In this project, we embarked on a journey to revolutionize pavement inspection and assessment through the integration of artificial intelligence (AI) and machine learning techniques. Our focus was on developing a model that not only accurately assesses pavement conditions but also predicts potential issues, thereby facilitating informed maintenance decisions.

Through the synergy of computer vision and machine learning algorithms, we devised a hybrid model that brings together the strengths of Convolutional Neural Networks (CNNs) and Gradient Boosting. This novel approach allowed us to extract intricate features from pavement images and combine them with quantitative measurements, leading to a comprehensive understanding of pavement conditions.

Our project's progress unveiled insights into both the promises and challenges of this pioneering approach. The utilization of CNNs for image feature extraction showed immense potential in capturing visual nuances that impact pavement conditions. The integration of Gradient Boosting facilitated robust predictions, equipping our model with the ability to offer valuable insights for future degradation.

However, our journey was not without hurdles. We navigated through the intricacies of class imbalance and the specter of overfitting, two challenges inherent in predictive modeling. The accuracy paradox highlighted the discrepancy between cross-validation success and test set performance, emphasizing the need for nuanced evaluation metrics like precision, recall, and F1-score. Furthermore, our exploration of overfitting underscored the importance of thorough validation on unseen data and the judicious application of regularization techniques.

In conclusion, our project bridges the chasm between traditional pavement inspection methods and the transformative potential of AI-driven approaches. While our results hold promise, we acknowledge that the path to accurate and robust predictions is a continuous evolution. Our hybrid model exemplifies the fusion of advanced technologies, paving the way for enhanced road safety and comfort. As we conclude this endeavor, we recognize that our work is a stepping stone in the ongoing journey to reimagine infrastructure maintenance and empower smart cities.

6 Reference

References

[1] Sudakov, O., Burnaev, E., & Koroteev, D. (2019). *Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks.* Computers & Geosciences, 127, 91-98.

[2] H. Weldegebriel, H. Liu, A. Haq, E. Bugingo, and D. Zhang, "A new hybrid convolutional neural network and eXtreme gradient boosting classifier for recognizing handwritten Ethiopian characters," *IEEE Access*, vol. 8, pp. 17804-17818, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2960161

[3] Q.-T. Bui, T.-Y. Chou, T.-V. Hoang, Y.-M. Fang, C.-Y. Mu, P.-H. Huang, V.-D. Pham, Q.-H. Nguyen, D.T.N. Anh, V.-M. Pham, and M. E. Meadows, "Gradient Boosting Machine and Object-Based CNN for Land Cover Classification," *Remote Sens.*, vol. 13, 2021, 2709. DOI: https://doi.org/10.3390/rs13142709

[4] M. Arifuzzaman, M. Aniq Gul, K. Khan, and S. M. Zakir Hossain, "Application of Artificial Intelligence (AI) for Sustainable Highway and Road System," *Symmetry*, vol. 13, 2021, 60. DOI: https://doi.org/10.3390/sym13010060

[5] Nhat-Duc Hoang, "An Artificial Intelligence Method for Asphalt Pavement Pothole Detection Using Least Squares Support Vector Machine and Neural Network with Steerable Filter-Based Feature Extraction," *Advances in Civil Engineering*, vol. 2018, Article ID 7419058, 12 pages, 2018. DOI: https://doi.org/10.1155/2018/7419058.

[6] Y. Wang and T. Ye, "Applications of Artificial Intelligence Enhanced Drones in Distress Pavement, Pothole Detection, and Healthcare Monitoring with Service Delivery," *Journal of Engineering*, vol. 2022, Article ID 7733196, 16 pages, 2022. DOI: https://doi.org/10.1155/2022/7733196.

[7] Gradient Boosting Trees for Classification: A Beginner's Guide — by Aratrika Pal — The Startup — Medium, https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2Fswlh%2Fgradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea

[8] Convolutional Neural Network with Extreme Gradient Boosting (CNN-XGBoost) Architecture. https://www.researchgate.net/figure/Convolutional-neural-network-with-extreme-gradient-boosting-CNN-XGBoost-architecture_fig2_350572008