Scour management in French rail infrastructure using machine learning

Tianyu Wang

Thèse présentée pour l'obtention du grade de

Docteur de l'Université Gustave Eiffel

Spécialité: Géotechnique

Ecole Doctorale: Sciences, Ingénierie et Environnement (SIE)

par

Tianyu Wang

Scour management in French rail infrastructure using machine learning

Thèse soutenue le 16/10/2023 devant le jury composé de:

Isam Shahrour	Université de Lille	Rapporteur
Pierre Breul	Université Clermont Auvergne	Rapporteur
Boulent Imam	University of Surrey	Examinateur
Maria Pina Limongelli	Politecnico di Milano	Examinatrice
Kenji Watanabe	University of Tokyo	Examinateur
Philippe Reiffsteck	Université Gustave Eiffel	Directeur de thèse
Chi-Wei Chen	SNCF Réseau	Encadrante
Franziska Schmidt	Université Gustave Eiffel	Encadrante
Christophe Chevalier	Université Gustave Eiffel	Invité

Acknowledgements

I would like to take this opportunity to express my gratitude to all those who have supported and guided me throughout my PhD.

Firstly, I would like to thank my supervisors Philippe Reiffsteck, Chi-Wei Chen, Christophe Chevalier, and Franziska Schmidt for giving me the opportunity of this PhD, and also for their guidance, support and encouragement during my study.

This PhD work is financed by the Société Nationale de Chemin de Fer (SNCF) and Association Nationale Recherche Technologie (ANRT). I'd like to acknowledge gratefully the engineers at SNCF who participated in the survey for the feature importance rank, and scour risk evaluations. I would also like to thank the technical support provided by the PLATIPUS project from DGII-OA SNCF Réseau regarding the web application construction. The help from Mr Julien Gabrielli, Mr Bastien Sage-Vallier and Mr Olivier Bougeard at SNCF is deeply acknowledged as well.

I'd also like to thank Mr Tsuyoshi Takayanaki at the Railway Technical Research Institute (RTRI), who provided the data of Japanese railway bridges. His expertise allows me to know well the situation in Japan and to do the comparison between French and Japanese models afterwards.

Then, I would like to thank all the jury members, starting from Prof. Isam Shahrour at Université de Lille, and Prof. Pierre Breul at Université Clermont Auvergne, who kindly accepted to be the reviewers and offered very insightful feedback on this thesis. I would also like to thank the examiners Dr. Boulent Imam at University of Surrey, Prof. Maria Pina Limongelli at Politecnico di Milano, and Prof. Kenji Watanabe at University of Tokyo.

I want to extend my thanks to my colleagues, Nader, Badr, Mara, Alessandra, Marwa, Thibault, Mathieu, for all the fun and difficult times we've had. I feel very lucky to know you. I hope all of you can finish the PhD successfully. I would like to thank Marie-Joe, who helps me a lot in administration during the PhD.

I'd also like to thank my friends, Dr. Jiayue Gu, Dr. Lisa Xiong, Dr. Zhehao Zhu, Dr. Kyeunghye Ahn, Prof. Feng Zhang, Yuechen Zhao, Qi Qiao. Some of them are in China but they are always there to support me, listen to my difficulties, and help me get through tough times.

Last but not least, my deepest appreciation is given to my parents for their unconditional love, encourage, and support during this PhD.

Abstract

Scour is a natural phenomenon originating from the erosive action of watercourse. It occurs on erodible riverbeds by carrying or excavating materials, and it can be particularly observed in the vicinity of obstacles such as bridge piers. Identifying rail infrastructure vulnerable to scour is one of the important tasks for transport agencies. In France, however, a practical process for identification has not existed yet for rail infrastructure. This thesis addresses the issue via machine learning based solutions, since they have shown good capacity in prediction recently. By using data from the French National Railway Company (SNCF), popular and classical algorithms are applied firstly to build the machine learning models. Then, the models are examined in terms of their robustness, and practicality. Meanwhile, they are compared with existing approaches such as ARPSA (Cerema, France) and scoring table (RTRI, Japan). Later, in order to understand the prediction results, the cutting-edge explainable artificial intelligence (XAI) methods are employed for model interpretation. Results from XAI are then compared with engineering expertise from SNCF. This thesis provides a quick yet accurate process for identifying rail infrastructure vulnerable to scour by benefiting from the novel data-driven approach. It helps improve the current inspection and maintenance process by emphasizing the importance of surrounding environmental factors. In the end, a user-friendly web application was built to ensure the accessibility of the research outcome.

Keywords: rail infrastructure; machine learning; scour risk; rail assets management

Résumé

L'affouillement est un phénomène naturel résultant de l'action érosive d'un cours d'eau. Il se produit sur les lits érodables en transportant des sédiments de la rivière en particulier au voisinage d'obstacles comme les piles de pont. L'identification des infrastructures ferroviaires vulnérables à l'affouillement est l'une des tâches importantes des gestionnaires de transport. En France, cependant, il n'existe pas encore de processus pour les infrastructures ferroviaires. Cette thèse aborde la question via des méthodes basées sur l'apprentissage automatique, puisqu'elles ont récemment fait preuve d'une bonne capacité de prédiction. En utilisant les données de la SNCF, des algorithmes classiques sont d'abord appliqués pour construire les modèles. Ensuite, les modèles sont examinés en termes de robustesse et de facilité d'utilisation. Enfin, ils sont comparés à des approches existantes telles que l'ARPSA (Cerema, France) et le « scoring table » (RTRI, Japon). Ensuiteles méthodes d'intelligence artificielle explicable (XAI) sont employées pour l'interprétation du modèle. Les résultats de l'XAI sont comparés à l'expertise des ingénieurs. Cette thèse propose un processus rapide mais précis d'identification des infrastructures ferroviaires vulnérables à l'affouillement. Elle permet d'améliorer le processus d'inspection et de maintenance en soulignant l'importance des facteurs environnementaux. Enfin, une application conviviale a été créée pour garantir l'accessibilité des résultats de la recherche.

Mots-clés: infrastructure ferroviaire; apprentissage automatique; risque d'affouillement; gestion des infrastructure ferroviaire

Publication lists

Journal papers

1. Wang T., Reiffsteck P., Chevalier C., Zhu Z., Chen C.W., Schmidt F. (2022) A novel extreme gradient boosting algorithm based model for predicting the scour risk around bridge piers: application to French railway bridges. *European Journal of Environmental and Civil Engineering*, 27(3), 1104–1121
   Location in thesis: Chapter 3 and Chapter 4

2. Wang T., Reiffsteck P., Chevalier C., Chen C.W., Schmidt F. (2023) An interpretable model for bridge scour risk assessment using explainable artificial intelligence (XAI) and engineers' expertise, *Structure and Infrastructure Engineering*, 1-13
   Location in thesis: Chapter 6

3. Wang T., Takayanagi T., Chen C.W., Reiffsteck P., Chevalier C., Schmidt F., Managing railway bridges crossing waterways through a machine learning based predictive maintenance policy, *Journal of Bridge Engineering* (under review)
   Location in thesis: Chapter 5 and Chapter 6

Confernce papers

1. Wang T., Reiffsteck P., Chevalier C., Chen C.W., Schmidt F., Application of

Random Forest algorithm in bridge scour risk prediction, Machine Learning Risk Assessment (MLRA2021), Wroclaw, Poland
Location in thesis: Chapter 4

2. Wang T., Reiffsteck P., Chevalier C., Chen C.W., Schmidt F., Maintenance Prédictive des Ouvrages d'Art avec des fondations en site aquatiques, 11ième Journées Nationales de Géotechnique et de Géologie de l'Ingénieur, Lyon 2022
Location in thesis: Chapter 4

3. Wang T., Reiffsteck P., Chevalier C., Chen C.W., Schmidt F., Machine learning based predictive maintenance policy for bridges crossing waterways, 9th Transport Research Arena (TRA), Lisbon, 2022
Location in thesis: Chapter 5

4. Wang T., Reiffsteck P., Chevalier C., Chen C.W., Schmidt F., An investigation into the robustness of machine learning models for bridge scour risk assessment, 4th International Symposium on Machine Learning and Big Data in Geoscience (ISMLG 2023), University College Cork, Ireland, 2023
Location in thesis: Chapter 5

5. Wang T., Takayanagi T., Reiffsteck P., Chevalier C., Chen C.W., Schmidt F., A comparison of the French and Japanese scour risk assessment procedures for railway infrastructure, The 11th International Conference on Scour and Erosion, Copenhagen, Denmark, 2023
Location in thesis: Chapter 5

Contents

List of Figures

List of Tables

Abbreviations

AI artificial intelligence

DP deep learning

FN false negative

FP false positive

ML machine learning

RF random forest

SHAP shapley additive explanations

SVM support vector machine

TN true negative

TP true positive

XAI explainable artificial intelligence

XGBoost extreme gradient boosting

Chapter 1

Introduction

1.1 Background of the study

A well-developed transport system is essential for a country's economic development. As one of the infrastructure elements, bridges are crucial points served for connection within the transport network, underpinning economic vitality, and logistics of communities (Sasidharan et al., 2021).

France currently operates the second-largest European rail network, with a total of 29,901 km of railway (UIC, 2022). For the French National Railway Company (Société Nationale des Chemins de fer Français, SNCF) and other transport agencies, managing and ensuring the safety of infrastructure under the extreme weather events become one of the priorities .

In most cases, scour is the physical process behind the collapse of structures after flooding. It is considered as a natural phenomenon originating from the erosive action of flowing watercourse and usually occurs on erodible beds by excavating or carrying away materials from the riverbed. Facing with the stochastic nature of hydrologic

events (e.g. flood), it is realized that scour related disasters are difficult to be avoided. In this case, a risk based evaluation model shall be necessary to complement the design and maintenance procedures. Tubaldi et al. (2017) established a probabilistic framework of scour depth by considering the hydrologic, hydraulic, and scour analyses using Markovmodel approach. Bento et al. (2020) proposed a risk-based methodology for scour risk at bridge foundations comprising hazards (extreme hydrological events), numerical modelling (river behaviour) using HEC-RAS and vulnerability (scour depth to foundation depth ratio) analyses. Besides the aforementioned research work, in practice, empirical based procedures were proposed to screen the high risk structures in a more rapid way such as the Design Manual for Roads and Bridges (DMRB) BD97/12 (British Highways Agency, 2012), EX2502 (HR Wallingford, 1992), ARPSA (Cerema, 2019) and the scoring table built by the Railway Technical Research Institute (RTRI) (Takayanagi et al., 2018).

At SNCF, Livret A (SNCF, 2005) was proposed for classifying the foundations of rail assets crossing or adjacent to waterways. This document, nonetheless, has been barely implemented in the field due to its complexity. As a result, scour risk assessment depends majorly on engineers' judgment and a risk level of bridge foundation is designated after the field inspection. Guidelines in other countries or sectors cannot be directly applied to the French rail infrastructure due to the data accessibility, different construction techniques, or geographical background.

Recently, the increasing availability of data, growing capabilities of hardware, and cloud-based solutions have all boosted a new type of maintenance policy called predictive maintenance (PdM). PdM is a method in which the service life of important parts is predicted based on inspection or diagnosis. Compared with reactive or preventive maintenance, PdM is condition-based and could decrease the total cost in the end (see Figure 1.1).

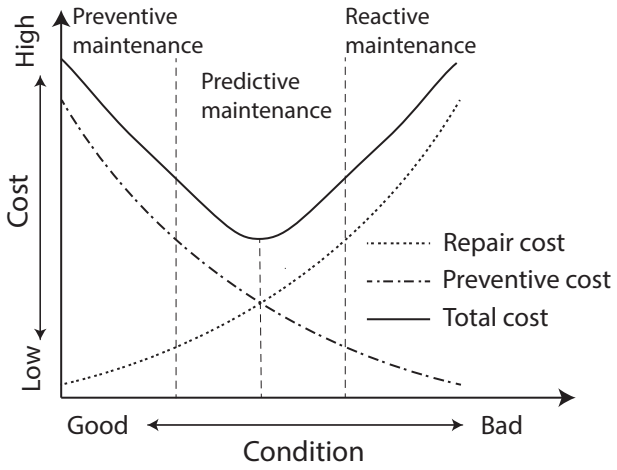


Figure 1.1: Three types of maintenance policies

Machine learning (ML) is often adopted in PdM for the definition of the actual condition of the system as well as for forecasting its future states (Susto et al., 2015; Yousefpour et al., 2021). Therefore, by realizing the limitations of existing methods, this novel data-driven approach is considered as another option since it can discover the patterns in data which is not apparent to humans. When looking at the current literature, ML has been adopted in several studies in civil engineering and shown encouraging results, such as bridge risk management (Cattan and Mohammadi, 2002; Elhag and Wang, 2007; Alipour et al., 2017), local scour depth prediction (Cheng and Cao, 2015), undrained shear strength (Mbarak et al., 2020), clay compressibility (Zhang et al., 2021b), liquefaction (Goh, 1994), etc.

1.2 Objectives

Global climate change is accelerating and the natural hazards such as flooding, drought will happen more frequently in the future. Therefore, having a simple yet effective procedure for managing the rail infrastructure regarding scour is indispensable for SNCF. The general objective of this PhD work is to propose a machine

learning based solution for evaluating the scour risk of rail infrastructure in France. The following objectives are set to achieve this goal:

- Review the typologies of infrastructure, maintenance policy at SNCF, and the practical guidelines adopted in other countries or transport agencies. Have the basic knowledge of machine learning and its application in geotechnical engineering;
- Collect information from inspection reports and construct the database that could be used for training the machine learning model;
- Build machine learning models with domain expertise. The proposed models should provide guidance to the engineers or inspectors who are seen as the end-users of the model and need to decide the following step after inspection. For example, whether the scour countermeasure is required for the infrastructure.
- If possible, propose an application which makes the research outcome easy to be used and understood by the engineers who don't know coding.

1.3 Overview of the research

This thesis consists of six chapters, which is outlined as follows:

Chapter 1 is the general introduction of the research work.

Chapter 2 presents the overview of the rail assets crossing waterways, including the surveillance and maintenance policy, construction techniques, and also different observed damages around the infrastructure. Then, basic knowledge of machine learning and its application in civil engineering is presented.

Chapter 3 introduces the process to build the datasets which are used to train the machine learning classifiers afterwards.

Chapter 4 introduces some popular machine learning algorithms. These algorithms are then used to build the machine learning model. The training and test results are presented in this chapter.

Chapter 5 compares the machine learning classifiers built in Chapter 4 from different perspectives. For example, the model performance is examined for the robustness, practicality. The machine learning models are also compared with existing methodologies used in Japan and France.

Chapter 6 explains one of the machine learning models that is constructed in Chapter 4. The cutting-edge explainable artificial intelligence (XAI) techniques are employed for interpretation. Furthermore, the research outcome of this study is demonstrated in a web application which could facilitate the use of rail engineers. The presentation of this web application is shown at the end of this chapter.

Chapter 7 is the general conclusions and perspectives of this work.

Chapter 2

State of the art

2.1 Preamble

The number of natural disasters has increased a lot over the past 50 years due to the climate change effect. Studies have shown that hydraulic events induced damages are the major causes for bridge failures. In 1978 in France, the collapse of the Wilson Bridge, which is a masonry bridge with 15 arches built in 1765-1778, affected nearly 100,000 people in the city of Tours.

This chapter introduces the relevant background information for the maintenance practical guidelines for scour risk assessment, basic concepts of machine learning and development of machine learning in civil engineering are reviewed.

2.2 Overview of historical rail infrastructure subject to scour

This section provides an overview of the railway assets subject to scour, including the difficulties for surveillance and maintenance, maintenance policy, foundation

construction techniques, and commonly observed damages in the field.

2.2.1 Specificities and difficulties

In the French rail network, except the infrastructure in high-speed rail lines (Train à Grande Vitesse, TGV), rail assets such as bridges, and retaining walls in other lines (e.g., Transport Express Régional, Intercités) were mostly constructed 120 years ago (SNCF, 2020). There are roughly 10,000 bridges crossing waterways and waterway retaining walls in the French rail network. Recently, the accelerating climate change bringing more severe and frequent natural hazards such as flooding, droughts, and heavy rains could undoubtedly pose a greater threat to the safety of infrastructure in the transport network (Nasr et al., 2021).

As a matter of fact, historical rail assets take a great proportion in Europe's rail network (Ozaeta García-Catalán and Martín-Caro, 2020). Each year, a huge amount of money is spent by the transport agencies for the reinforcement or reconstruction of rail infrastructure affected by floods. In France, the annual cost to regenerate the bridges subject to scour or other hydraulic events is roughly 4 million euros, according to SNCF (2019). The Austrian Federal Railways company (ÖBB) estimated over 100 million euros losses due to floods in recent years (Kellermann et al., 2016). Passenger travel disruptions due to floods were estimated to cost up to GBP 60 million for the UK railway network, and the indirect losses (e.g., impact on economic productivity) could be 10 times larger (Lamb et al., 2019).

The aforementioned data indicates the necessity to have an effective monitoring and maintenance policy. A proper policy for managing infrastructure could not only ensure the safety of rail network but also decrease the high cost related to repair work after the occurrence of scour disaster. However, there are several challenges and difficulties in monitoring and maintaining the aging rail infrastructure in the rail system.

Firstly, bridge management involves numerous data collection and data analysis techniques. Most railway bridges served in regional express lines (e.g., TER, Intercités) are between 75 and 170 years old. Some of them were partly destroyed during World War I or World War II and reconstructed after. Therefore, the design drawings could be incomplete or partly lost, and most of the drawings have not been digitized yet at SNCF. Besides, there could be the riverbed evolution, and flow path change since construction. In order to understand well the bridge and its surrounding environment, detailed inspections (underwater inspection through diving) are necessary. Sometimes, in-situ tests such as geotechnical surveys or impact vibration tests (Nishimura and Tanamura, 1989) are conducted as well. All of them are at a high cost and sometimes require even additional human resources to proceed.

The second challenge originates from the mode of monitoring. The most efficient and cost-effective method to deal with scour is to monitor the evolution overtime and to program scour countermeasure work accordingly (Prendergast and Gavin, 2014). One of the commonly used techniques for bridge monitoring is visual inspection. When the depth of watercourse is deep or the velocity is high, an underwater foundation inspection through diver is employed (Sasidharan et al., 2021). However, such visual inspection is only conducted during the low-flow periods and cannot be realized in flooding when the scour risk is the highest, due to safety reasons. Moreover, the issue is exacerbated as the scour hole may be filled in when floodwater subsides and get unnoticed during visual inspection, which could mislead the real extent of the problem (Sasidharan et al., 2021).

The third challenge comes from the external threat. The unequivocal extreme climate events have been recorded more often in the past decade. Study showed that the intensity and frequency of floods are possible to increase due to global warming (Few, 2003). Compared with new constructions, historical bridges are more vulnerable when facing natural hazards resulting from material degradation. Besides the

changing climate, another threat comes from the increased loading conditions due to the growing demand of transport on rail. If judging by the current design procedures, the initial design loads have more or less been passed, leading to the growing vulnerability of rail assets.

2.2.2 Surveillance and maintenance policy

In order to guarantee the safety of rail assets, especially for which in an ageing system, a proper maintenance policy is important. Figure 2.1 illustrates the inspection–maintenance process at SNCF. The maintenance work starts with a field inspection. The railway engineers must make a decision regarding the action to be done afterwards based on site observations. The most optimistic case is that the bridge foundation is at a very low risk level and the next inspection is scheduled in no more than six years. In other cases, corresponding scour countermeasures are going to be scheduled, and there are three types of cases: immediate regeneration, regeneration, and preventive work. According to the status of infrastructure, acceptable time to finish the corresponding maintenance work is different (varying from less than one year to six years). In some extreme cases, the severe damages (bridge collapse, deformation) could result in totally disrupted circulation. The scour countermeasures and/or the reinforcement of bridge foundation should be done immediately.

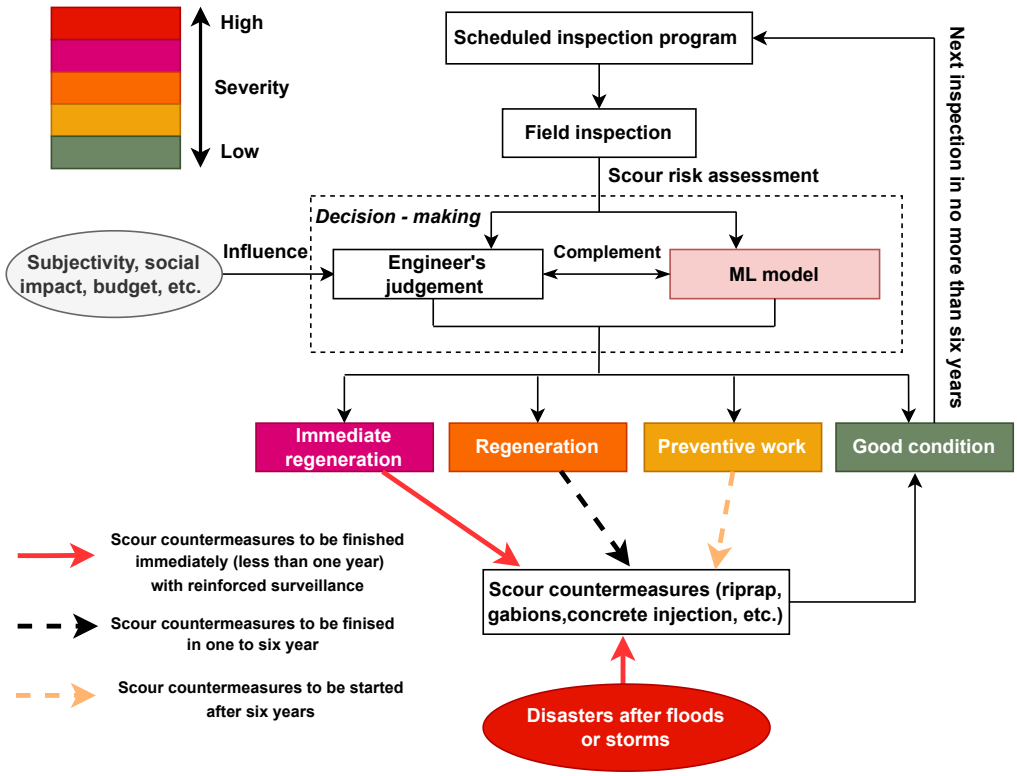


Figure 2.1: Surveillance and maintenance policy for rail infrastructure subject to scour.

The railway engineer makes decisions primarily based on their domain knowledge and expertise. It can be seen from Figure 2.1 that this step is also influenced by other aspects, such as his/her subjectivity affected by working experience, budget constraint, and social impact (importance of the bridge and rail line in the whole transport network). Therefore, the objective of this PhD work is to propose a machine learning model, which could serve as a complement for engineer's judgment and optimize the decision-making process in the end.

2.2.3 Types of foundations

This subsection presents briefly the types of foundations for historical railway bridges crossing waterways.

2.2.3.1 Timber pile foundation

Timber pile foundation has existed since Roman and it was employed until the end of the 19th century. Before the naissance of caisson foundation, it was the only way to construct on compressible soil, such as silt, clay, and clayey sand.

Normally, the spacing between piles is often close to 1 m and the diameter of the pile varies from 0.20 m to 0.35 m. The length of the pile is around 10 m but can reach 20 m in some rare cases by using a joint. Figure 2.2 shows the schema of a bridge foundation constructed on timber piles.

Generally speaking, timber pile foundation construction consists the following steps:

1. driving the piles into the soil to make sure they could support the loads;
2. cutting up the piles in the same plane and connecting the top together;
3. covering a raft on the top of the piles to continue the following work.

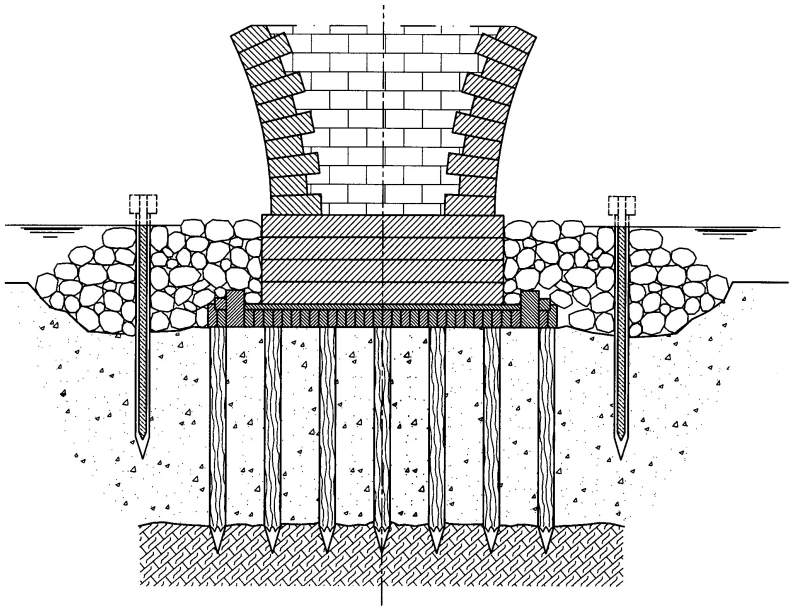


Figure 2.2: Schematic presentation of a timber pile foundation (SNCF, 2005).

2.2.3.2 Caisson

Caisson is a commonly used technique to work on the foundations of a bridge pier. Water is pumped out in the caisson to keep the work environment dry. There exist several construction techniques for caisson foundation. In general, the compressed air allows limiting the infiltration of water inside the caisson.

Figure 2.3 shows a pneumatic caisson which is closed at the top and open at the bottom. The compressed air forces the water out and keeps the working environment dry. This process creates an airtight space where workers can excavate mud and rock debris until hitting the bedrock. Concrete will be poured afterwards to form a solid bridge pier.

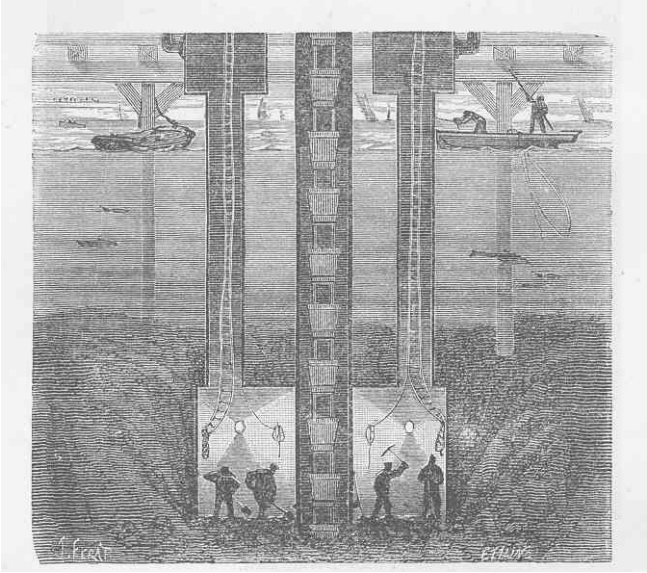


Figure 2.3: Pneumatic caisson (Sornel, 1872).

2.2.3.3 Mass concrete

When the soil layer which supports the loads is not too far from the water level, the foundation can be constructed on mass concrete. This kind of foundation is realized by using underwater concrete or pumping out water and then pouring concrete. A

cofferdam, which is often made by timber piles, sheet piles or earthfill dam, surrounds the foundation allowing creating a dry working environment (see Figure 2.4).

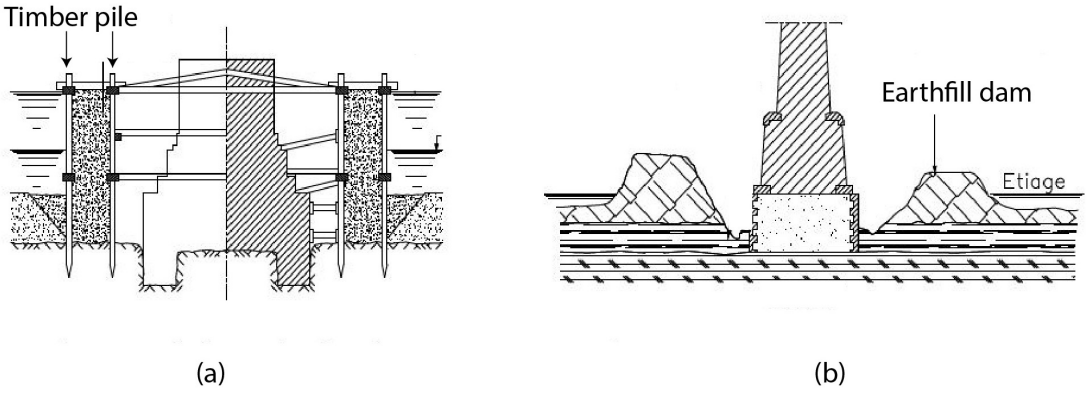


Figure 2.4: Mass concrete foundation surrounded by: (a) timber pile; (b) earthfill dam (adapted from SNCF, 2005).

2.2.4 Types of damages

This subsection describes and classifies the commonly seen damages during field inspection. The asset and its surrounding environment are decomposed into four parts: riverbank, foundation(infrastructure), channel, and superstructure. Damages described in this subsection are normally observed from visual inspections. Special inspections (e.g., underwater foundation inspection through diving, timber pile foundation inspection) are required if the river depth is profound, velocity is high or the foundation is constructed on timber piles.

2.2.4.1 Riverbank

The common observations (damages) around riverbank are listed as follows and shown in Figure 2.5:

- landslide;

- erosion;
- excessive vegetation;
- contraction of flow due to debris.

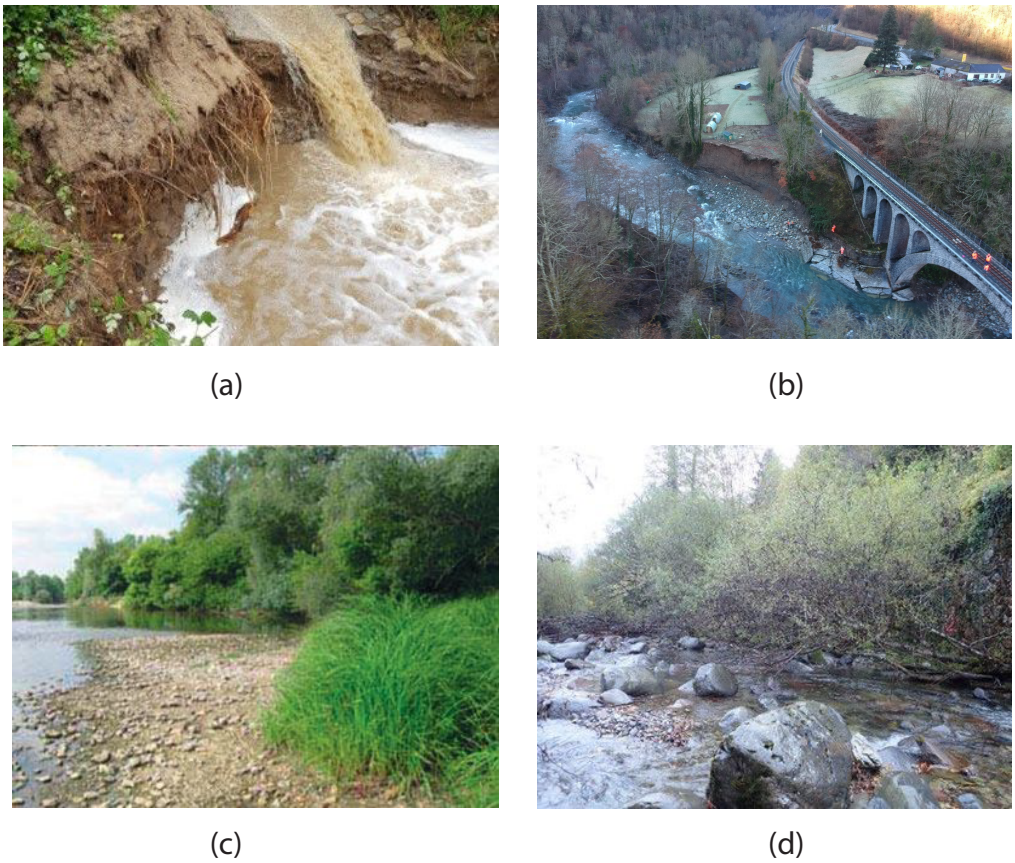


Figure 2.5: Observations (damages) around riverbank: (a) landslide; (b) erosion; (c) excessive vegetation; (d) contraction of flow due to debris (SNCF, 2020)

2.2.4.2 Foundation (infrastructure)

The causes for foundation damages are due to the degradation of the structural element or the interaction with watercourse. The types of damages are classified as follows (Ozaeta García-Catalán and Martín-Caro, 2020).

Damages due to the structural element are:

- irregular cracks on foundation;
- abraded and rotten timber piles;
- corrosion of caisson;
- loss of scour protection.

Damages related to the interaction of watercourse are:

- general scour;
- local scour around foundation.

Figure 2.6 shows photos of several commonly seen damages around bridge foundation.

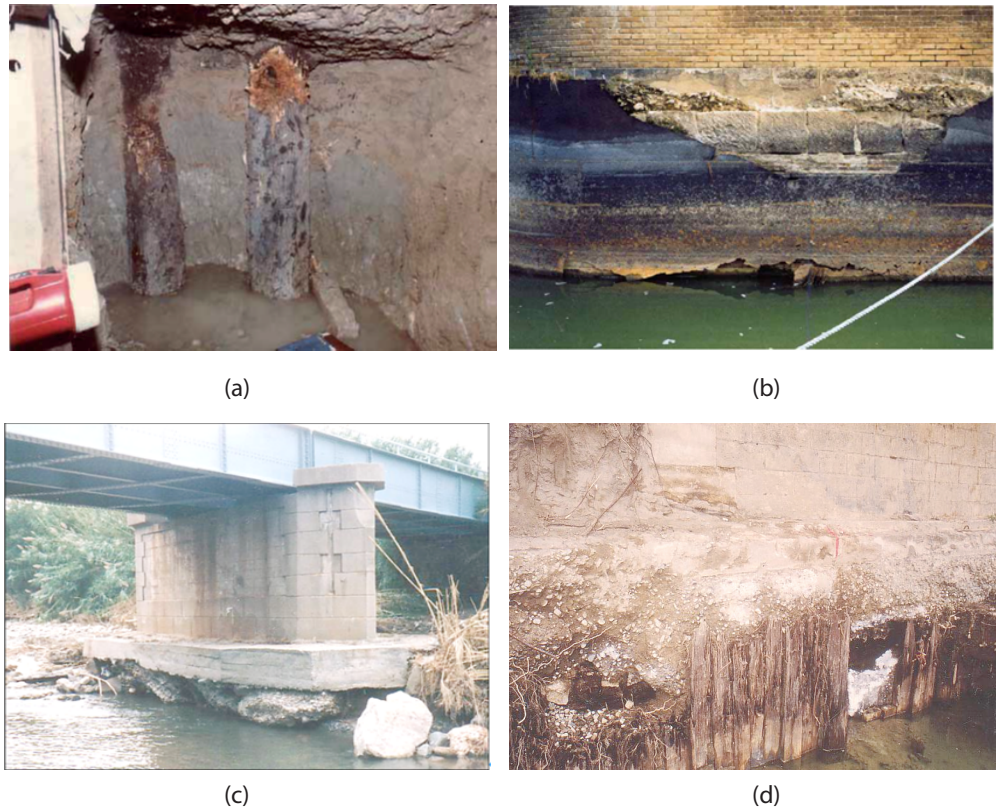


Figure 2.6: Photos of damages on foundation: (a) rotten timber piles; (b) corrosion of caisson; (c) local scour; (d) loss of scour protection (SNCF, 2005).

2.2.4.3 River channel

Damages on river channel are categorized as follows and some of them are illustrated in Figure 2.7.

- general scour (lowering of riverbed);
- debris;
- bars (due to the sediment deposition);

- riverbed movement;
- excessive vegetation.

Figure 2.7: Photos of damages on river channel: (a) debris; (b) bars (sediment deposition); (c) excessive vegetation.

2.2.4.4 Superstructure

Although damages on superstructure are beyond the scope of foundation inspection, but they could indicate the instability of foundation (Ozaeta García-Catalán and Martín-Caro, 2020). Such damages may be:

- cracks in longitudinal, vertical or transverse direction;

- mechanical failure of masonry.

Figure 2.8 presents the commonly observed damages on superstructure.

(a)

(b)

Figure 2.8: Photos of damages on superstructure: (a) mechanical failure of masonry; (b) longitudinal crack.

2.3 Basic knowledge of scour

2.3.1 Background

Scour is the leading cause for bridge failures (Deng and Cai, 2010; Pizarro et al., 2020; Dikanski et al., 2018). It is considered as a natural phenomenon originating from the erosive action of flowing watercourse and usually occurs on erodible beds by excavating or carrying away materials from the riverbed.

According to Shirole and Holt (1991), more than 60% of bridge failures were related to scour or other hydraulic effects between 1950 and 1991 in the United States. Since 1840, more than 100 collapses regarding railway bridges were due to scour and caused 15 fatalities in the United Kingdom (Van Leeuwen and Lamb, 2014).

2.3.2 Types of scour

The types of scour typically encountered are presented in this subsection.

- **General scour** : general scour is also called as natural scour. Melville and Coleman (2000) define general scour as "that scour occurring irrespective of the presence of any human-imposed structure". It occurs due to the natural variability of river stream flows and sediment regime, considering the influence from the catchment to the river scale. Riverbed degradation, bend, contraction scour are all part of general scour (Pizarro et al., 2020).

- **Contraction scour** : compared with local scour, contraction scour has received much less attention in recent research (Ghazvinei et al., 2014). Contraction scour occurs when riverbed width decreases and watercourse velocity increases. With contraction scour, riverbed material can be mostly or all removed due to the increased flow velocities and shear stresses. Conditions result in contraction scour may be : (1) the constriction of watercourse (natural contractions or

bridge contractions), (2) natural berms along the riverbanks due to sediment deposits, (3) ice formations, (4) debris, (5) vegetation in the channel (Ekuje, 2018).

- **Local scour** : local scour happens when there is an structural obstruction splitting (e.g., piers) or disrupt (e.g., abutments) the flow. According to Benn (2013), local scour at bridge pier is considered as the most predominant type of scour. Factors related to local scour include but not limited to intensity of flow, sediment size of riverbed material, flow depth, attack angle, shape, width and length of pier and abutment (Ekuje, 2018).

Figure 2.9 illustrates the aforementioned different types of scour that could happen near a bridge pier.

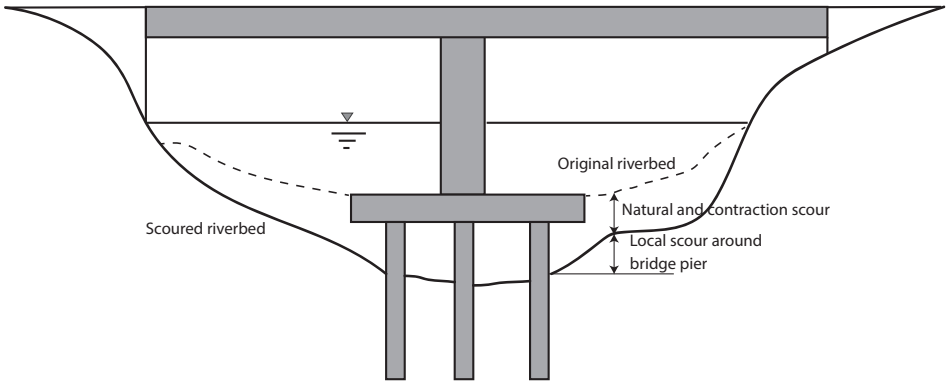


Figure 2.9: Schematic presentation of different types of scour.

2.3.3 Climate change impact on scour

As mentioned in the previous subsection, bridge scour is the foremost cause of bridge failures worldwide (Dikanski et al., 2018). It removes and excavates the material from

riverbed due to the erosive action of watercourse. The removal of material starts when the erosive capacity of water flow exceeds its ability to resist motion, thus the scour process commences (Annandale, 2006). Furthermore, Annandale (2006) points out that scour depth increases when the erosive capacity of water flow increases, which in return lifts and drags the sediment at the riverbed by the flow velocity. Such process could consequently lower the riverbed in the main channel (e.g., general and contraction scour) and around bridge piers and abutment (e.g., local scour).

It is broadly acknowledged in the literature that climate change due to anthropogenic emissions of green house gases (e.g., CO_2) is taking place (IPCC, 2013). The latest report of the Intergovernmental Panel on Climate Change (IPCC) indicates that western and central Europe will have more pluvial and river flooding as a result of the global average temperature increase (Economist, 2021). Watts et al. (2015) asserts that *"anthropogenic climate change would expectedly modify rainfall, temperature, and catchment hydrological processes around the world resulting in challenging water-related adaptations".*

The increased frequency of flood events has an adverse impact on the life-cycle performance of affected railway infrastructure, since the design code is on the basis of past events and assume the load will be the same in the future (Mondoro et al., 2018), which leads to growing vulnerability of these infrastructure under future hazards.

To summarize, the increase in river flood discharge resulting from extreme precipitation due to climate change effect could increase scour risk and ultimately, endanger the safety of infrastructure and cause more failures.

2.4 Practical procedures for scour risk assessment

In order to prevent scour disasters, transport agencies normally conduct regular visual inspections to evaluate the risk of infrastructure. Practical guidelines are pro-

posed to help field engineers assess the risk in case they don't have enough knowledge in the domain. Before proposing a procedure for the rail assets in France, we look at firstly the existing approaches in Japan, the United Kingdom, and France in rail or road sectors, since they share similar construction techniques, or geographical background to French assets. After reviewing these approaches, the advantages and disadvantages of each method are compared and discussed at the end of this section.

2.4.1 Rail sector

2.4.1.1 Scoring table in Japan

Proposed by the Railway Technical Research Institute (RTRI), the Japan Railways Groups (JR) employs a scoring table (Takayanagi et al., 2018) to assess the scour risk of railway bridges.

In this approach, factors related to scour risk are divided into three categories : environmental condition of river, structural conditions of bridge pier, and protection conditions of bridge pier. The evaluation items in each category are considered having an important impact on scour risk assessment and they are chosen on the basis of past disasters happened in Japan.

It should be noted that many parameters used in empirical formulas for calculating scour depth (Laursen, E. M. and Toch, A., 1956; Froehlich, 1988; Melville, 1997; Melville and Sutherland, 1988; Hager and Unger, 2010; Hong et al., 2012; Link et al., 2017; Melville and Yee-Meng, 1999; Pizarro et al., 2017; Yanmaz and Altinbilek, 1991) are reflected in this method. Moreover, types of scour countermeasure as well as the damage level of scour countermeasures have been taken into account in this approach.

RTRI's scoring table is shown from Table 2.1 to Table 2.3. The importance of each evaluation item to final scour risk is quantified by the score shown in the column

Score.

Table 2.1: RTRI scoring table-Environmental condition of river

Evaluation item	**Choice**	**Score**
Topographical land form	Plain	10
	Valley plain	10
	Alluvial fan	0
	Mountainous area	5
Constriction of river width	Absent	15
	Present	0
Riverbed material	Sand	10
	Gravel	0
	Exposed rock or boulder	10
Overall riverbed degradation	Present	0
	Absent	10

Table 2.2: RTRI scoring table - Structural conditions of bridge pier

Evaluation item		**Choice**	**Score**
Bridge pier location relative to river bend		Straight river or inside of river bend	15
		Outside of river bend	0
Bridge pier location relative to floodplain		In river flow	5
		Floodplain without revetment	10
		Floodplain with revetment	25
		Floodplain without revetment and adjacent to river flow	0
		Floodplain with revetment and adjacent to river flow	15
Downstream drop structure		Absent	20
	Height	Up to 1m	5
		1 - 2m	0
		More than 2m	0
	Deterioration	Present	*
	Construction range	Only a part of river course	*
Relative embedment depth		Spread foundation or pile foundation	Figure 2.10
		Caisson foundation	
Variation of embedment depth		Increase or decrease in depth by more than 1.5m in comparison with that in the previous inspection	*
Bedrock contact		Absent	0
		Probable	15
		Present	30

Table 2.3: RTRI scoring table - Protection conditions of bridge pier

Evaluation item		Choice	Score
Absent			0
Unknown			0
Deterioration of basket foundation		Present	0
		Absent	5
		Unknown	0
Block footing protection	Deterioration	Absent	20
		Partial	5
		Overall or washed away	*
		Unknown	0
	Connection	Present	5
Expansion of footing	Relative embedment depth	Top of protection work < Riverbed	20
		Bottom of protection work < Riverbed < Top of protection work	10
		Riverbed < Bottom of protection work	*
	Deterioration	Present	*
		Unknown	0

Detailed explanations of several evaluation items (e.g.,"*Constriction of rived width*", "*Bridge pier location relative to rived bend*") may refer to the work of Takayanagi et al. (2018). The score of the evaluation item "*Relative embedment depth*" shown in Table 2.2 is calculated based on Figure 2.10. The embedment depth is calculated using equation 2.1

$$Relative\ embedment\ depth = L/B \tag{2.1}$$

where L is the bridge foundation embedment depth (m) and B is the bridge pier width (m).

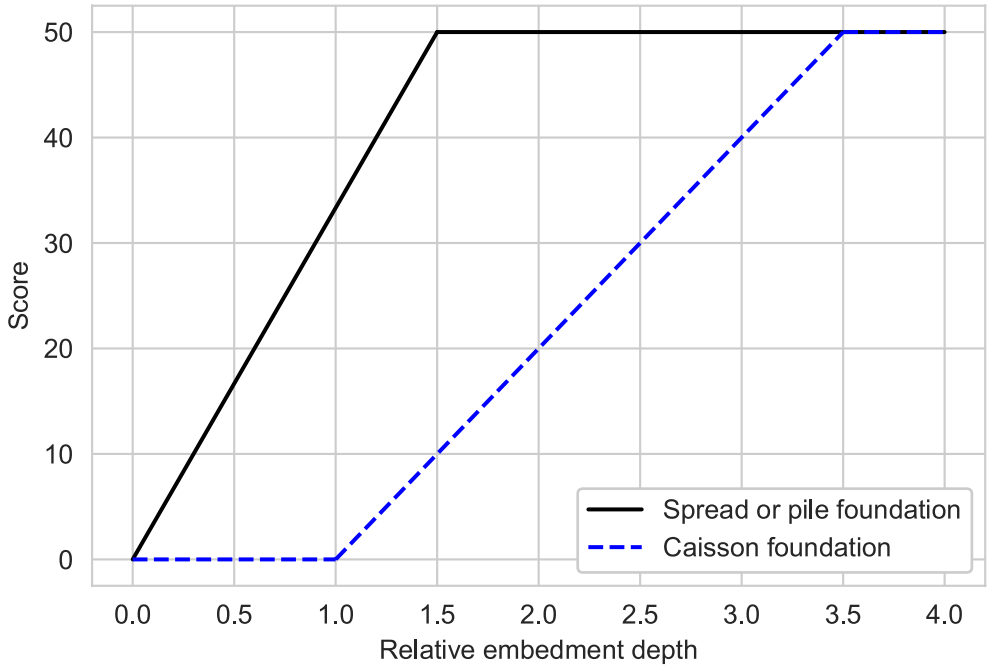


Figure 2.10: Relationship between relative embedment depth and score (Takayanagi et al., 2018).

The final bridge score is calculated by summing the score of evaluation items shown from Table 2.1 to Table 2.3. If the final score is less than 110, a detailed inspection and more complex studies shall then be needed. It's noteworthy that several choices in evaluation items with a "*" mark are considered strongly related to a potential scouring disaster. Therefore, if one of these choices are included in the investigated bridge pier, it is regarded as high risk and needs more detailed inspection immediately, regardless of the sum of scores.

The score of each evaluation item and the threshold value (110) were calibrated on historic survey data of 77 bridge piers in Japan. The same 77 bridge piers were considered as high risk by railway engineers without using scoring table. In addition, it has been tested and confirmed that the results from scoring table and railway engineers are generally in agreement with each other (Takayanagi et al., 2019).

2.4.1.2 EX2502 in the United Kingdom

EX2502 (HR Wallingford, 1992) is the standard for the assessment of scour for railway bridges over water in the UK. The procedure for evaluating the scour depth in EX2502 involves calculating general scour, local scour and bend scour (if any).

The total scour is calculated as the sum of local scour and general scour, as shown in equation (2.2)

$$d_t = d_l + d_g \tag{2.2}$$

where $d_t = total\ scour$; $d_l = local\ scour$; $d_g = general\ scour$.

To estimate local scour and general scour depth, important features are such as river bend, relative flow depth, angle of attack, pier shape, sediment size and debris blockage (if any) (Ekuje, 2018). All equations used in EX2502 for predicting the general and local scour depth are based on small-scale laboratory experiments with limited field data for verification (López et al., 2014; Arneson et al., 2012).

The total scour depth d_t should sometimes be revised by subtracting an adjustment factor (AF), as indicated in equation (2.3).

$$d_t = d_t - AF \tag{2.3}$$

AF is actually the difference in bed level at a cross section upstream of the bridge and the bed level from where the foundation depth was measured. It should be noted that AF can be applied only when there is enough information indicated in the divers' reports or information from monitoring equipment like sonar. In other cases, AF should be taken as zero.

A preliminary priority rating (PPR) can then be calculated by using the total scour depth and the foundation depth

$$PPR = 15 + ln\left\{\left(d_t - d_f\right)/d_f + 1\right\} \tag{2.4}$$

where $d_t = total\ scour\ depth$; $d_f = foundation\ depth$. The PPR can also be obtained by using the graph on Figure 2.11.

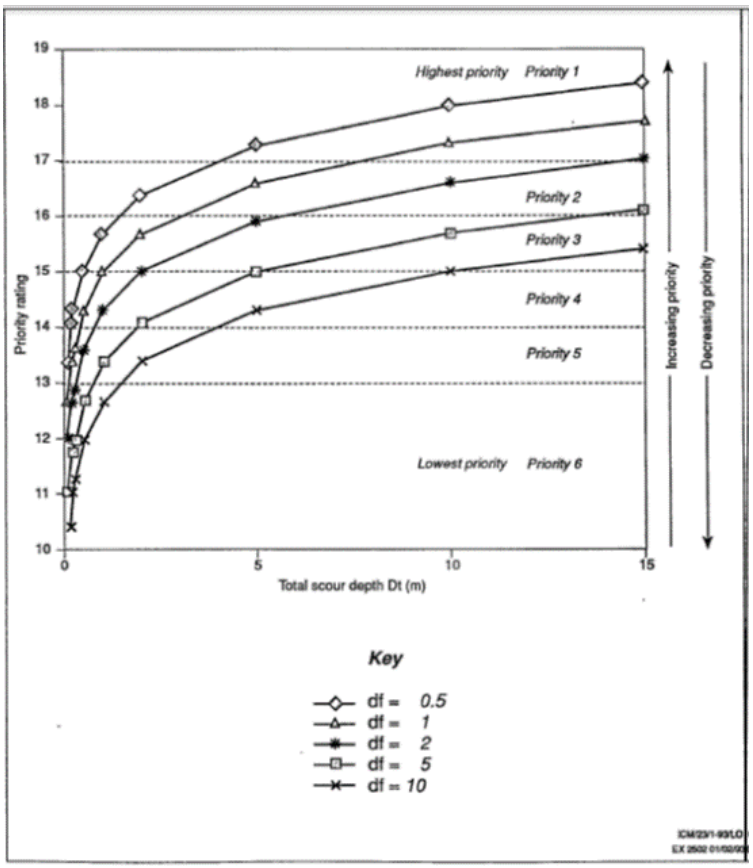


Figure 2.11: Final priority rating versus total scour depth (source: HR Wallingford, 1992)

The final priority rating (FPR) is calculated by considering a number of additional features which may influence of failure such as river stability and river type (TR). TR is calculated by using the equation (2.5)

$$TR = \frac{1}{17}(sum\ of\ channel\ stability,\ bank\ stability,\ channel\ slope) \tag{2.5}$$

and the final priority rating (FPR) is presented in equation (2.6).

$$FPR = PPR - TR \tag{2.6}$$

In the end, the risk of scour can be determined from Table 2.4.

Table 2.4: Final priority rating categorisation

Priority rating	**Category**	**Priority category**
>17	1	High
16 - 17	2	High
15 - 16	3	Medium
14 - 15	4	Medium
13 - 14	5	Low
<13	6	Low

One of the limitations of EX2502 is that it does not respond to changes in river discharges induced by climate change. The river depth is advised to be calculated as:

$$Y_u = 0.185 W_u^{0.7} \tag{2.7}$$

where $Y_u = flow\ depth\ in\ the\ channel$;
$W_u = channel\ width\ upstream\ of\ the\ bridge$.
It is obvious that the flow depth does not respond to changes in river discharge. Therefore, Manning's equation is suggested to estimate the flow depth (British Highways Agency, 2012). Ekuje (2018) presented the implement of Manning's equation for making the flow depth responsive to changes in discharge for improvement.

Another limitation of EX2502 is that it is not based on flow velocity and the influence of flow intensity on local scour (e.g., ratio of velocity due to discharge to critical velocity (V/V_c)) is not considered. Therefore, it cannot be used to calculate critical discharge based on critical velocity, where the maximum local scour occurs (Ekuje, 2018).

Detailed process for local scour (d_l), general scour(d_g) and additional factors (TR) is in presented in HR Wallingford (1992).

2.4.2 Road sector

2.4.2.1 ARPSA in France

ARPSA (Cerema, 2019) is the acronym for "Risk anlysis for Structures in scourable site" (in French : Analyse de Risque des Ponts en Site Affouillable). Proposed by Cerema, it is a French guideline to evaluate the risk of road bridges to scour. Like the other practical procedures, it is a risk-based guideline which could help asset managers to anticipate the future risk and manage infrastructure. It should be noted that ARPSA was developed and calibrated for road bridges. Extra work should be done to make it adaptable for French railway bridges (Takayanagi et al., 2019).

ARPSA comprises 3 steps' analyses: a pre-filtering step, a simplified analysis step, and a detailed analysis if necessary. Hazard, vulnerability and consequence are studied separately in ARPSA.

Figure 2.12 shows the scoring table for hazard factors in step 1 (pre-filtering) analysis. Generally speaking, the scour hazard level takes into account the general scour (A1), contraction scour (A2) and local scour (A3). Scour hazard characterization is mainly influenced by river morphology and hydrodynamic factors (Takayanagi et al., 2019). The scour hazard level in step 1 is then defined by three levels shown in Table 2.5. The hazard level in step 2 refers to a simplified quantitative analysis. It calculates scour depth by using classic equations (Arneson et al., 2012). Parameters used in step 2 to calculate scour depth are, among others, flow, flow depth, pier width, riverbed sediment diameter.

Influencing factors for scour hazards		**Score**			
River flow conditions	Fluvial rivers	2	A11		
	Torrential rivers	3.5			
	Mountain torrential rivers and rivers under cyclonic conditions	5			
Type of riverbed sediment	Rocky showing substratum	0	A12	0	A34
	Blocks	1		0.4	
	Gravels, pebbles…	1.6			
	Cohesive soils (silts, clays)	2.8		1	
	Sand	3.5			
General scour	2 x A11 x A12 - 5		A1		
Contraction scour	Low impact on hydraulic section due to the structure	0.5	A2		
	Reduction from 15 to 40% of hydraulic section due to the structure	2			
	Reduction > 40% of hydraulic section due to the structure or when structure is under pressure flow	6			
Dimension of support considered as flow obstacle	No projecting support or pile or abutment	0	A31		
	Width ≤ 2m	1.5			
	2m < width ≤ 4m	2.5			
	Width > 4m	3.5			
Pier shape	Favorable configuration: circular or square or no obliquity against flow direction	1	A32		
	Obolong, or with obliquity versus flow direction	3			
	Other cases	2			
Riverbed evolution	Stable	1.1	A33		
	Presence of dunes H>1m or alluvial mobile sandbank	1.3			
Local scour	1.2 x A31 x A32 x A33 x A34 x A11		A3		
Hazard level (total scour)	**A1 + A2 + A3**		**A**		

Figure 2.12: ARPSA scoring table : hazard factors (step 1) (Durand et al., 2019).

Table 2.5: ARPSA- Hazard (A) level classification (step 1)

Hazard level	**Note**
Low	A < 3
Medium	$3 \leq A < 7$
High	A ≥ 7

Figure 2.13 is the scoring table for vulnerability factors in step 1 analysis. Here, the vulnerability level is a combination of three subdomains: foundation vulnerability (V1), support sensibility to foundation destabilization (V2) and bridge deck vulnerability to different movements (V3) (e.g., total or differential settlements of the supports, swaying). The vulnerability level (V) is then classified for each bridge (see Table 2.6 for step 1 classification).

ARPSA also includes a table for consequence (ISE). It is based on factors related to bridge economical or technical significance within the transport network, patrimonial value, and consequences on serviceability, potential consequences of structure failure, etc.

Influencing factors for scour vulnerability		**Score**	
Construction periods	After 1976	-1	V11
	1951-1975	3	
	Before 1950	5	
Type of foundation	Deep or half deep foundations	2	V12
	Spread <wide> foundations	5	
	Other cases or lack of data	10	
Surveillance	Recent inspection with no scouring process	0	V13
	Other cases	4	
Foundation vulnerability	- If no projecting pile or abutment: 0 - Otherwise: V11+V12+V13		V1
Material	Concrete or metal, in a good condition	1	V21
	Masonry, or damaged concrete or metal	3	
Shape of the obstacle	Sharp cutwater	0	V22
	Cylindrical support or circular cutwater	0.5	
	Rectangular support	1	
Support sensibility to foundation destabilization	V21 + V22		V2
Type of structure	Concrete culvert of truss-bridge	1	V3
	Other cases	2	
Desk sensibility to relative movements, swaying or partial loosening of the support	V3		V3
Vulnerability level	- **If V1 ≤ 5: V1** - **If V1 > 5: V1+V2+V3**		**V**

Figure 2.13: ARPSA scoring table : vulnerability factors (step 1) (Durand et al., 2019).

Table 2.6: ARPSA- Vulnerability (V) level classification (step 1)

Vulnerability level	**Note**
Low	$V \leq 8$
Medium	$8 < V \leq 12$
High	$V > 12$

Bridge scour risk is determined at first through criticality (C) level by referring to hazard (A) and vulnerability (V) levels. Later, the scour risk is a combination of consequence (ISE) level (see Table 2.7) and criticality (C). Table 2.8 and Table 2.9 show the two matrices to determine risk level in step 1 analysis.

Table 2.7: ARPSA- Consequence (ISE) level classification

Consequence level	**Note**
Very low	$0 \leq ISE < 4$
Low	$4 \leq ISE < 8$
Middle	$8 \leq ISE < 12$
High	$12 \leq ISE < 16$
Very high	$16 \leq ISE \leq 20$

Table 2.8: ARPSA - Criticality (C) level classification (step 1)

Criticality	**Low vulnerability**	**Medium vulnerability**	**High vulnerability**
Low hazard	Low criticality	Low criticality	Medium criticality
Medium hazard	Low criticality	Medium criticality	High criticality
High hazard	Medium criticality	High criticality	High criticality

Table 2.9: ARPSA - Risk matrix (step 1)

Risk	**Low criticality**	**Medium criticality**	**High criticality**
Very low consequence	Very low risk	Very low risk	Low risk
Low consequence	Very low risk	Low risk	Medium risk
Medium consequence	Low risk	Medium risk	High risk
High consequence	Medium risk	High risk	Very high risk
Very high consequence	High risk	Very high risk	Very high risk

After step 1, bridge at medium, high or very high risk requires a step 2 semi-quantitative analysis. Similarly to step 1, it calculates the hazard, and vulnerability levels.

Hazard level in step 2 is the function of total scour depth P, which is the sum of general scour depth P_1, contraction scour depth P_1 and local scour depth P_3. Equations to calculate P_1, P_2 and P_3 are respectively from the work of Ramette (1981), Laursen (1963),Arneson et al. (2012).

The vulnerability level in step 2 is determined with similar criteria in step 1 but a French standard for the classification of structures (IQOA) is added. Besides, the foundation tilting due to potential local scour is taken into account.

ARPSA was tested and calibrated on 83 road bridges including 11 in torrential and 4 in cyclonic conditions (Takayanagi et al., 2019). Detailed information on ARPSA scoring table is presented in Cerema (2019) and Durand et al. (2019).

2.4.3 Comparison among the procedures

The procedures presented before are commonly used guidelines for scour risk evaluation in practice. Table 2.10 and Table 2.11 list the advantages and disadvantages of each procedure respectively.

Table 2.10: Advantages of scoring table, EX2502 and ARPSA

Procedure	**Advantages**
Scoring table	1. A very practical and simple procedure to be adopted in practice. 2. Each bridge element (pier) is assigned a risk level. 3. Types of damages over scour countermeasures (e.g., sheet piling, gabion) are considered in risk evaluation.
EX2502	1. Formulas for estimating local scour, and general scour depth are established on published scour equations. 2. Conditions of riverbanks are included to assess the stability of the river. 3. Each bridge element (pier) is assigned a risk level.
ARPSA	1. Besides vulnerability and hazard, consequence factors are included in risk evaluation. 2. Bridge superstructure is considered for scour risk evaluation.

Table 2.11: Disadvantages of scoring table, EX2502 and ARPSA

Procedure	**Disadvantages**
Scoring table	1. History of the bridge, the stability of riverbank and river channel are not taken into account. 2. Consequence factors are not included in the procedure. 3. The score of each evaluation item is based on Japanese morphological background. 4. River flow, which could be changed due to climate change effect, is not taken into consideration. 5. Bridge abutments, and retaining walls cannot be evaluated by scoring table.
EX2502	1. Consequence factors are not included in the procedure. 2. Formulae are calibrated based on small-scale laboratory experiments with limited field data. 3. River depth, and local scour can not be responsive to the change of river flow or velocity.
ARPSA	1. The threshold value for determining the vulnerability, hazard and consequence levels are based on engineering expertise and several case studies.

Among all three approaches, scoring table proposed by RTRI is the easiest one to be applied in practice. The parameters required in this approach are much less than the others. However, considering scoring table is established based on Japanese geographical background, it may be not exactly adaptable to the French situation.

EX2502 is established based on published and very popular scour equations. Compared with scoring table, it requires many field measurements (e.g. river width, flow depth upstream or under the bridge). A quantitative analysis can then be conducted.

Having these measurements could sometimes be expensive and time-consuming in practice. Although equations for simulating the river depth are proposed in EX2502, they are not responsive to the change of river flow. The local scour depth estimated in EX2502 is not related to velocity as well.

ARPSA is the only one among the presented procedures considering the consequence aspect for scour risk evaluation. Since it is a procedure for highway bridges, some evaluation items are not adaptable to railway bridges.

In the end, each procedure has its own advantages and disadvantages, and none of them can be used directly to railway bridges in France. The objective of this study is to propose a novel procedure for infrastructure in the French rail network. In this research, the proposed ML models are compared with scoring table and ARPSA (see section 5.4 in Chapter 5), since data required in these approaches are similar to the current maintenance policy at SNCF.

2.5 The fundamentals of machine learning

This section reviews the fundamental concepts of machine learning. It clarifies the definition of artificial intelligence, machine learning, deep learning and introduces different types of machine learning. Training and testing principle, model evaluation methods are also presented in this section.

2.5.1 Artificial intelligence, machine learning and deep learning

According to the Cambridge Dictionary, artificial intelligence (AI) is *"the science of making computers do things that human beings can do."* In other words, it mimics and displays "human" cognitive skills by machines, especially computer system. Specific applications of AI include but not limited to natural language processing, computer

vision, speech recognition, self-driving cars.

Machine learning is a branch of AI. According to the definition of IBM, machine learning focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Figure 2.14 illustrates the process to build a spam filter using traditional programming and ML technique. Traditional programming needs to write a detection algorithm in advance with key words (such as "credit card", "good news", and "free") and then flag emails if one or several of words are detected. The program needs to be tested and repeated several times until it's good. ML approach, on the other hand, learns automatically words or phrases that are good predictors for spam from the given spam examples, even the unusual frequent patterns can be learnt. This program is much shorter, easier to be maintained, and most likely more accurate (Aurélien Géron, 2019).

Deep learning is a subset of ML. It can be generated as a sophisticated and mathematically complex evolution of machine learning algorithms. It consists of multiple layers of interconnected nodes, each building upon the previous layer to optimize the prediction or categorization (Zhang et al., 2022).

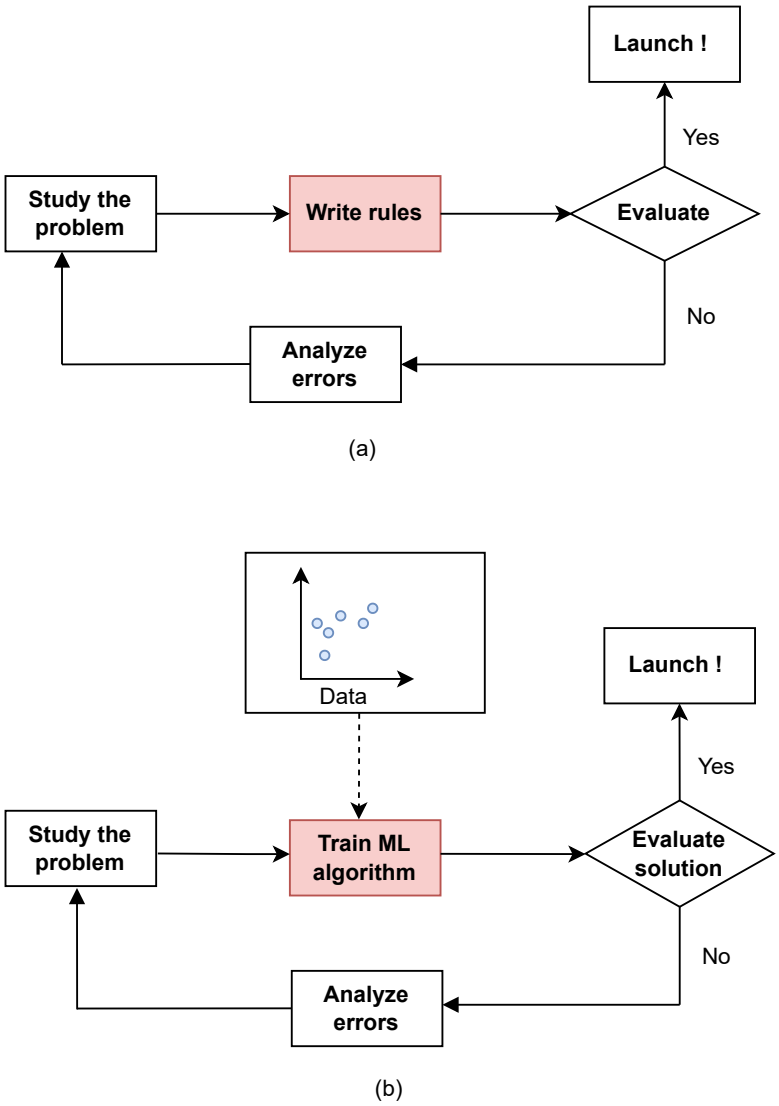


Figure 2.14: Spam filter construction through traditional programming (a) and ML approach (b) (Aurélien Géron, 2019).

Figure 2.15 presents the scope of AI, ML and deep learning. To conclude, ML and deep learning are both types of AI. Deep learning is a subset of ML which uses artificial neural networks for learning.

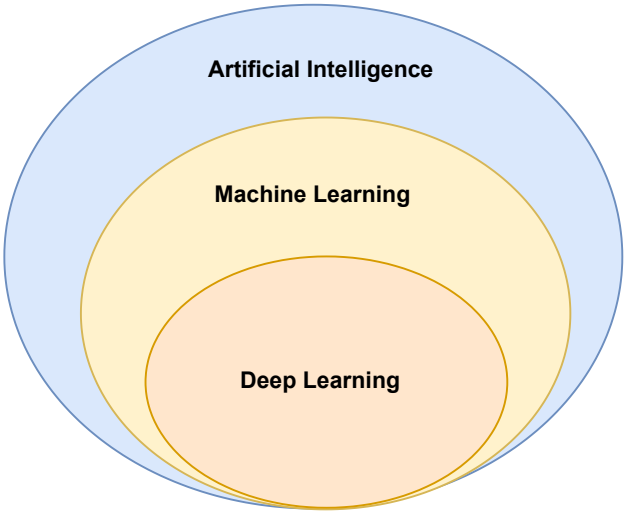


Figure 2.15: Artificial intelligence, machine learning and deep learning.

2.5.2 Types of machine learning

There exist four categories of ML, namely supervised learning, unsupervised learning, semi-supervised learning, as well as reinforcement learning.

2.5.2.1 Supervised learning

Supervised learning must use data including the desired solution, which is called *labels*. The labeled dataset is then fed to train algorithms and to classify data or predict outcomes accurately (see Figure 2.16). The model adjusts its weights until it has been fitted appropriately using training data. Below are some commonly used supervised learning algorithms:

- logistic regression;
- support vector machine (SVM);
- k-nearest neighbors;

- random forest (RF);
- extreme gradient boosting (XGBoost);
- neural networks.

Among the listed algorithms, SVM, RF, XGBoost and neural networks are selected to train the ML model because they've shown promising results in other related studies. Chapter 4 presents more detaily these algorithms.

Supervised learning has been widely used in geotechnical engineering, including soil properties (Bejarbaneh et al., 2018), tunneling (Zhou et al., 2017), landslides (Moosavi et al., 2014), remote sensing (Yousefpour et al., 2021). A review of ML applications in civil engineering is presented in section 2.6.

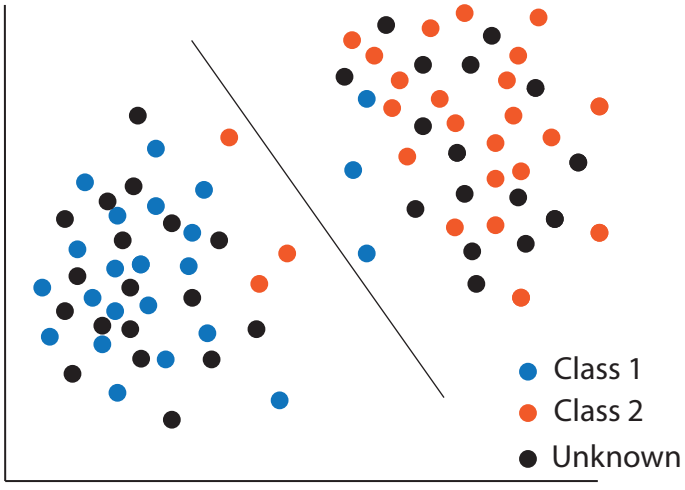


Figure 2.16: A labeled dataset for supervised learning.

2.5.2.2 Unsupervised learning

Unsupervised learning uses ML algorithms to analyze and cluster unlabeled dataset, as shown in Figure 2.17. Compared with supervised learning, the algorithms can dis-

cover the underlying patterns without prerequisite knowledge (human intervention).

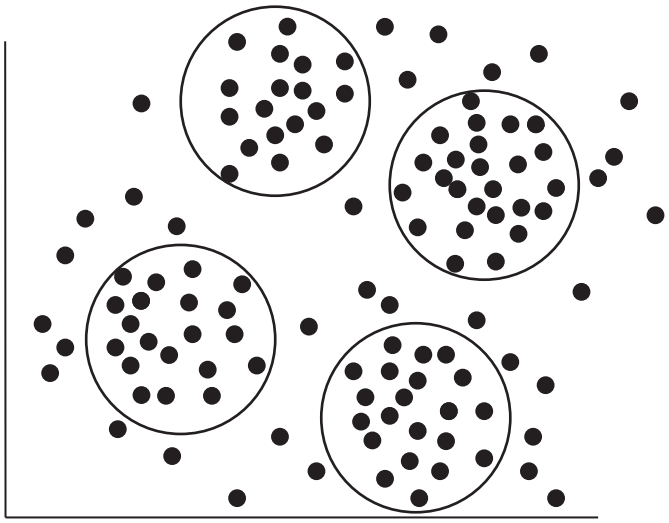


Figure 2.17: An unlabeled dataset for unsupervised learning.

Unsupervised learning is often used to clustering, which means to group similar data points together based on their characteristics. It can also be used for dimensionality reduction, in which the goal is to simplify the data without losing so much information. Another important task for unsupervised learning is *association rule learning*. It digs into large amount of data and discovers relationships between features. The *association rule learning* is often used in marketing analysis for decision-making.

Commonly used unsupervised learning algorithms are:

- k-Means;
- principal component analysis (PCA);
- kernal PCA.

2.5.2.3 Semi-supervised learning

Semi-supervised learning falls between supervised and unsupervised learning. During the training phase, it uses only a small amount of labeled data set to guide classification and feature extraction from a larger, unlabeled dataset. Semi-supervised learning is often used in image recognition for searching photos.

2.5.2.4 Reinforcement learning

Reinforcement learning is based on rewarding desired behaviours and/or punishing undesired ones. In general, the learning system (also called *agent*) can interpret environment, take actions and learn through trial and error.

Considering the objective of this study (scour risk evaluation) and number of variables that we are gonna have, supervised learning is chosen to train the model. Therefore, the datasets must include factors that have impacts on scour risk (input parameters) and the corresponding risk (output). Detailed work regarding dataset preparation is shown in Chapter 3.

2.5.3 Model training and testing

2.5.3.1 Data division

ML model is capable of making predictions by learning from data. Typically, the collected data is divided into two parts, namely training and test sets.

Training set is used to fit the model (e.g., weights and biases in a neural network). The ML model is trained on the training set. The size of the training set is much larger than the other sets (validation or test sets) because it is hoped that the model can learn as much as possible from it.

Test set can provide an unbiased evaluation of a final model fitted on the training set. It is only used when the model training phase is finished. In other words, the

ML model cannot be tuned any further after testing set.

In some cases, validation set is being needed to provide an unbiased evaluation of a model fit on training set while fine-tuning the hyperparameters (e.g., number of hidden layer or number of iterations in a neural network). It can be served for model regularization through early-stopping to avoid overfitting.

2.5.3.2 Underfitting and overfitting

Overfitting and underfitting are two crucial concepts in ML and are the causes for the poor performance of ML model in most cases. Figure 2.18 shows the examples of underfitting, optimum learning and overfitting.

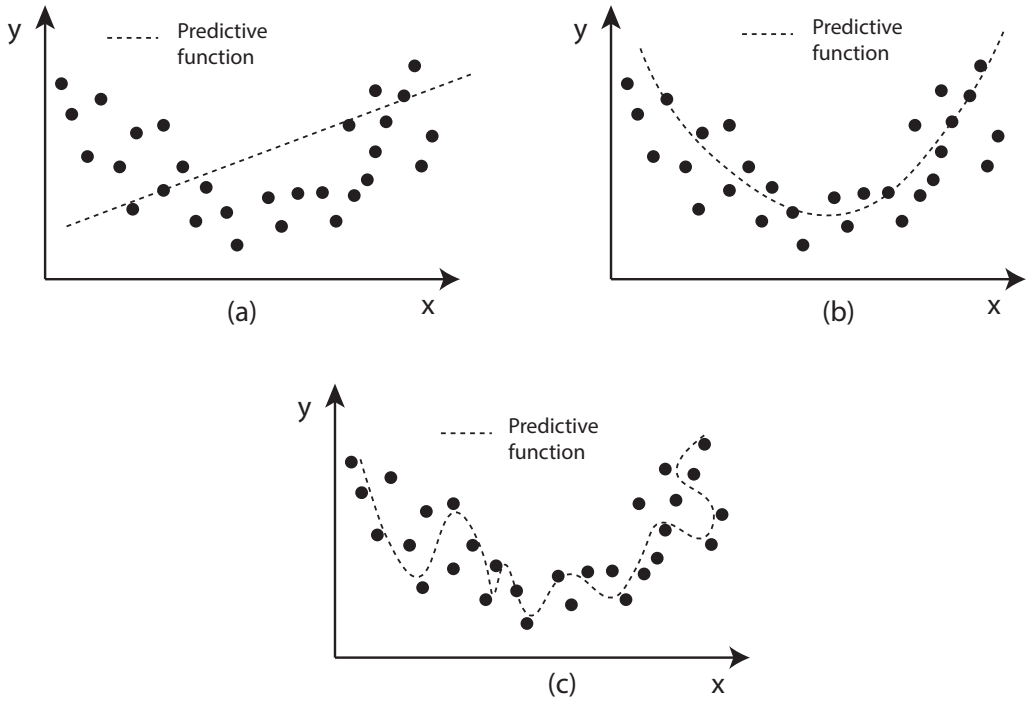


Figure 2.18: Examples of underfitting (a), optimum (b), and overfitting (c).

As shown in Figure 2.18 (a), underfitting occurs when the model is too simple and

it cannot represent the underlying structure of data. For example, a linear regression model is used for life satisfaction prediction. Obviously, it is acknowledged that the reality is more complex than a linear model.

Some techniques to avoid underfitting are:

- increasing model complexity;
- increasing the size of data, number of variables;
- removing noise from the data;
- increasing the number of epochs for training.

Overfitting (see Figure 2.18(c)) is the opposite of underfitting. It happens when the model fits exactly against the training data but not generalizes well. In other words, the model predicts well the training data but not the test. The reasons for overfitting are whether the model training time is long or the model is too complex, so the model starts to learn from the noise data or irrelevant information.

In order to avoid overfitting, possible solutions are:

- collecting more training data;
- early stopping to pause training;
- feature selection to identify the most important parameters;
- regularization (e.g., Lasso regularization);
- ensemble learning by using a set of classifiers.

As a matter of fact, there is no specific rule to define the threshold value for overfitting or underfitting. Most documents indicate overfitting is "*when the model*

performs well well on the training data but does not on test data". Similarly, it is difficult to say how much data (or variables) is needed for a machine learning model to avoid underfitting or overfitting. The decision can only be made once we know the training and test results. Therefore, without training and testing, it's difficult to know in advance whether the collected data or selected variables are enough to have an accurate classifier or not.

2.5.3.3 Evaluation measurements

In order to know how well the model learns from the training data, it needs to be tested in an unseen dataset, which means the test set as mentioned in section 2.5.3.1. If the model performs well on training set but poorly on test set, it indicates overfitting. If both training and test sets error are small, it means the model learns well from the data.

The model performance evaluation measurements are chosen based on the types of problems. If the model output is a numerical value (e.g., soil compression index, foundation settlement), typical performance measurements are the root mean square error (RMSE), coefficient of determination (R^2), mean absolute percentage error (MAE), etc.

By contrast, if the model prediction is classes (e.g., soil type, risk class), the evaluation measurement should be changed correspondingly.

One of the most common ways to evaluate a classifier is to use a use a confusion matrix, as shown in Table 2.12. There could be four possible outcomes when comparing the predicted values with actual values, namely true positive (TP), false positive (FP), false negative (FN), and true negative (TN).

Table 2.12: Confusion matrix

		Actual values	
		Positive	Negative
Predicted values	Positive	True Positive (TP)	False positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Commonly used evaluation measurements can be calculated from this matrix and are shown from equation 2.8 to equation 2.11. It should be noted that in a perfect classifier, the number of FN and FP cases are 0. In other words, the ideal values for accuracy, precision and recall are equal to 1 whereas 0 for false positive rate.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.8}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.10}$$

$$False\ positive\ rate\ (FPR) = \frac{FP}{FP + TN} \tag{2.11}$$

Accuracy (equation 2.8) is one of the most popular measurements to evaluate a classifier. It's calculated as the percentage of correctly classified cases among all test data. Simply using accuracy score cannot perfectly evaluate the model performance. For example, achieving 90 percent accuracy is trivial for an imbalanced dataset. In other words, if the proportion of one class is extremely superior to another, high accuracy score does not indicate that the classifier has a satisfactory performance.

To address this issue, precision (equation 2.9) is employed to calculate the per-

centage of true positive samples among all predicted positive samples. At the same time, recall (equation 2.10) calculates the percentage of true positive samples among all actual positive samples. In other words, a high precision means that the model has a high probability to give a correct prediction for positive classes, and a high recall indicates that the classifier is capable of correctly detecting the positive classes. False positive rate (equation 2.11) is the counter part of recall (Liu et al., 2021). It's the fraction of false positive cases among all actual negative samples, namely the probability of false alarm.

Area Under the ROC Curve (AUC) is an alternative measurement to evaluate a classifier based on a Receiver Operating Characteristic (ROC) curve (Fawcett, 2006). In ROC curve the true positive rate (recall) is plotted against the FPR at different thresholds [1] as shown in Figure 2.19.

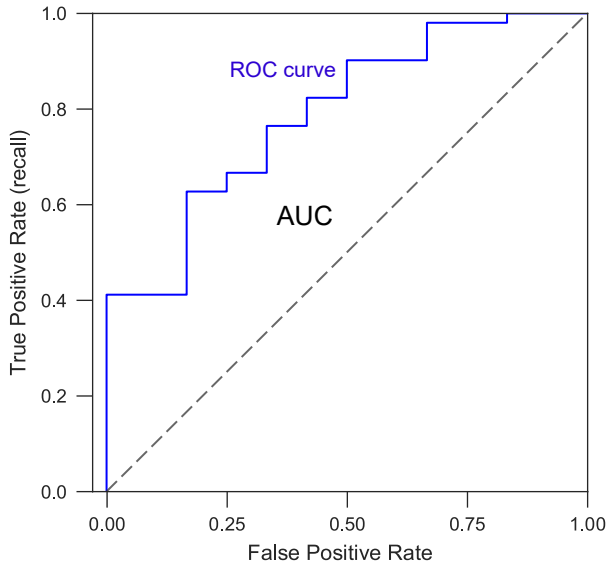


Figure 2.19: Example of an ROC curve.

[1] In a binary classification problem, the algorithm returns to a probability which ranges from 0 to 1. The probability allows mapping to a binary category. By default, the threshold value is 0.5. If the probability is larger than 0.5, the sample belongs to class A. On the contrary, if the probability is less than 0.5, the sample could be in class B.

The dotted line represents the ROC curve of a random classifier. As indicated by its name, the area under ROC curve is AUC. AUC measures how well the classifier distinguishes between two classes. A random classifier will have an AUC equal to 0.5 whereas a perfect classifier will have an AUC equal to 1.

2.6 Machine learning development in civil engineering

2.6.1 Model development process

As a branch of AI, ML model is designed to use data and algorithms to mimic human learning process. It should be noted that in the literature, there hasn't existed a standardized procedure to study a specific problem via AI/ML (Naser, 2021). Nonetheless, Shahin (2013) pointed out the procedures and directions to be systematically investigated (see Figure 2.20) when developing a machine learning model.

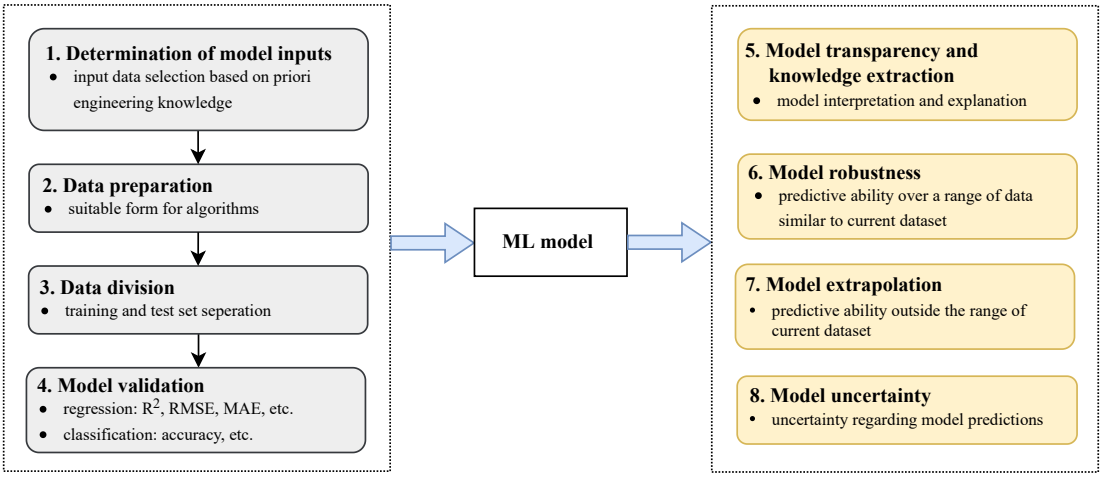


Figure 2.20: Methodology for ML model development (adapted from Shahin, 2013).

Figure 2.20 illustrates the process to develop a ML model. It can be seen that determination of model input, data preparation, data division and model validation

(point 1 to 4) are primary steps to construct a ML model for prediction. Once the model is trained, to enhance its reliability and practicability, point 5 to 8 are needed.

This thesis follows the methodology shown in Figure 2.20. It not only focuses on ML model construction, but also on checking model robustness for performance evaluation, and incorporating prior knowledge for model transparency. Chapter 3 presents the work regarding the selection of input parameters, data preparation and data division. The machine learning models are validated by different measurements in Chapter 4. Chapter 5 examines model robustness and Chapter 6 interprets the model for its transparency.

2.6.2 Examples of applications

With large volumes of observation and monitoring data available for researchers in civil engineering, ML based solutions are being integrated into various subdomains. When looking at the literature, in most studies the objective for applying machine learning is prediction. However, several studies also try to interpret the ML model besides prediction. Some research work with these purposes are shown below.

2.6.2.1 Prediction

- **Domain of scour**

Lots of research work has applied machine learning algorithms to predict the bridge pier scour depth. Compared with empirical formulas, ML approach does not require predefined coefficients to determine the relationship among key parameters. Among all machine learning models, artificial neural networks (ANNs) have shown encouraging results for local scour depth prediction (Cheng and Cao, 2015; Najafzadeh et al., 2013; Zounemat-Kermani et al., 2009; Hosseini et al., 2018; Lee et al., 2007; Tola et al., 2023). Bateni et al. (2007) used bayesian neural networks for predicting the equilibrium and time-dependent scour depth around bridge piers. Five key parame-

ters, namely, flow depth, mean velocity, critical velocity, mean grain diameter, and pier diameter are used for prediction. Results in this paper showed that ML models predict more accurately than the existing mathematical expressions. Toth and Brandimarte (2011) applied ANN models by using both field and laboratory observations for predicting the maximum scour depth around bridge piers. An external validation dataset is employed for evaluating the ANNs and literature formulae. Results confirm that the ANNs outperform the conventional approaches.

Despite the encouraging results, data in above studies is from laboratory tests. ML has rarely been applied in the engineering practice due to the need for a large amount of data required for training and validation. This PhD work addresses the issue by using data from SNCF.

- **Other problems in civil engineering**

Besides scour, AI and ML have recently been applied in other contexts. For example, ML models have been adopted for soil properties prediction, such as undrained shear strength (Mbarak et al., 2020; Zhang et al., 2021b), compression index (Park and Lee, 2011), and clay compressibility (Zhang et al., 2021a). Apart from this, liquefaction (Goh, 1994; Liu et al., 2006; Pal, 2006), landslide (Dahigamuwa et al., 2016; Li et al., 2012; Liu et al., 2021; Pradhan and Lee, 2010), pile capacity (Pham et al., 2017; Ghorbani et al., 2018; Harandizadeh et al., 2019), pile settlement (Saeedi Azizkandi et al., 2014; Armaghani et al., 2018), stability of underground space (Adoko et al., Adoko et al.; Ghasemi and Gholizadeh, 2019; Mahdevari et al., 2013), and tunneling in terms of tunnel boring machine (TBM) performance (Naghadehi et al., 2018; Elbaz et al., 2019; Mahdevari et al., 2014) are all very popular topics to use ML and AI technics.

ML has also been used for bridge maintenance and risk management. Cattan and Mohammadi (2002) used neural networks to predict subjective ratings for bridge

conditions given by experts. It was found that by giving the right input variables, neural network had a much better performance compared with conventional statistical methods and fuzzy-logic approach. Elhag and Wang (2007) established ANN in bridge risk score and risk categories assessment. Alipour et al. (2017) applied decision trees and random forest algorithms to predict the load-capacity rating of bridges, which was rated only by engineers' judgement before. A user-friendly software (App) was developed by Abedi and Naser (2021) to identify fire-vulnerable bridges.

2.6.2.2 Model interpretation via explainable artificial intelligence (XAI)

Besides using for prediction, explainable artificial intelligence (XAI) has been employed to understand the physical phenomenon through explaining the black-box ML model. XAI has currently been adopted in structural engineering, but rarely in geotechnical engineering. Shahin (2013) mentioned the importance of developing an explainable ML model in civil engineering. Naser (2021) discussed XAI and potential implications of XAI from a structural engineering perspective.

SHAP (SHapley Additive exPlanation) is a very common and popular model for interpretation in XAI. Bakouregui et al. (2021) used SHAP model to identify the contribution of input parameters for determining the load-carrying capacity of reinforced concrete columns. Mangalathu et al. (2020) established an explainable ML model by applying SHAP for understanding the failure mode of reinforced concrete columns and shear walls. Wang et al. (2023), Wakjira et al. (2021), Hu et al. (2021), Ma et al. (2023) have also applied SHAP model in their studies.

Based on these applications, the ML models are built at first for prediction (Chapter 4). They are interpreted via SHAP model and other XAI approaches afterwards (Chapter 6).

2.7 Summary

More frequent and severe flood events will happen in the future due to climate change. Enhancing the resilience of rail infrastructure under extreme weather conditions will be a key challenge for transport agencies.

In this literature review chapter, the challenges and difficulties for managing and maintaining the rail infrastructure have been introduced firstly. The aging infrastructure, the ancient construction techniques, and also a diverse typologies for damages around infrastructure could all bring challenges for asset managers.

Later, by revealing the physical model of scour, it is clearly realized that the climate change will inevitably accelerate the scour process and make an adverse impact to the life-cycle performance of railway infrastructure.

Then, the practical guidelines from rail and road sectors, namely scoring table, EX2502 and ARPSA are presented as examples on how to assess the scour risk in an effective way in practice. The advantages and disadvantages of each method are discussed.

It has been realized that in order to help field engineers, the assessment procedures must be simple enough to be applied in a large amount of assets. However, the challenge could be that when dealing with such amount of assets from a national level, many factors need to be taken into account: history, construction technique, surrounding environment, surveillance frequency, etc. Nevertheless, scour is complicated phenomenon which refers to the interaction of subsoil-foundation-watercourse and the underlying mechanism has not been completely understood yet. Current approaches cannot be directly used for the rail infrastructure in France.

To tackle the aforementioned issues, machine learning based solutions will be proposed in this study, since it can discover the patterns in data which is not apparent

to human. The last two sections of this chapter review the fundamentals of machine learning, including types of learning, definitions of underfitting and overfitting, and model performance evaluation methods. Later, the development of machine learning in civil engineering is presented. The methodology introduced here is followed in the rest of the study.

Chapter 3

Data preparation

3.1 Introduction

In order to build a machine learning model, having a proper dataset is the first step. This chapter introduces the process for dataset preparation.

In general, there are four types of elements affected by scour: bridge pier, bridge abutment, bridge wing wall (see Figure 3.1) and retaining wall (see Figure 3.2). When looking at the current literature, most research work focuses on studying scour at the bridge pier through physical modelings (Melville and Sutherland, 1988; Melville and Yee-Meng, 1999; Sumer et al., 2007; Amini et al., 2012) or numerical approaches (Richardson and Panchang, 1998; Zhu and Liu, 2012; Lu et al., 2008; Foti and Sabia, 2011). Scour at bridge abutment is often studied separately from bridge pier and the formulas for calculating local scour depth at bridge pier and abutment are not the same (Arneson et al., 2012).

Like the previous studies, we separated the raw data as two datasets, one for bridge pier and another for bridge abutment, bridge wing wall and retaining wall (noted as

Abutment&Wall hereafter). The two datasets have gone through the same process for feature selection and data preprocessing.

This chapter is organized as follows: Section 3.2 introduces the collection of raw data. Later, how to select input variables is shown in Section 3.3. The necessary approaches for making the data ready to be used in the ML model are presented in Section 3.4. In the end, the final two datasets as well as their visualizations are shown in Section 3.5.

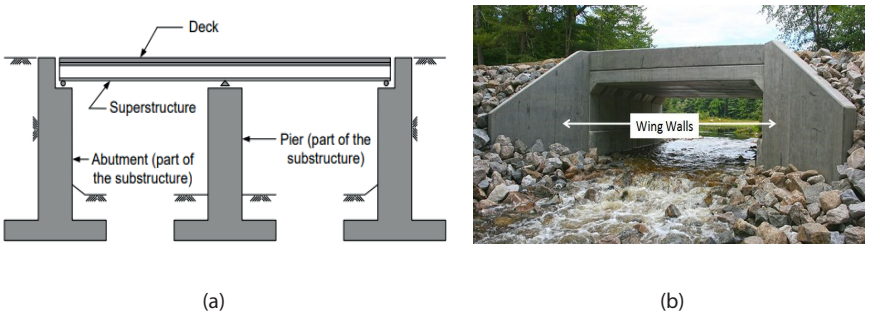


Figure 3.1: Components of a bridge: (a) pier and abutment; (b) wing wall (source: www.hpdconsult.com/parts-of-a-bridge).

Figure 3.2: Example of a retaining wall.

3.2 Data collection

Figure 3.3 illustrates the 19 SNCF regional offices. In general, it's the regional office who programs and is in charge of the surveillance, inspection and maintenance of the railway infrastructure. In order to build a database in a short period but which is representative enough, 16 out of 19 regions' bridges and walls are included in our original database, as shown in Figure 3.3. A wide range of topographical and hydraulic conditions is covered : watercourse varies from streams in mountainous area to main rivers in urban area such as the Seine River, the Loire River, etc.

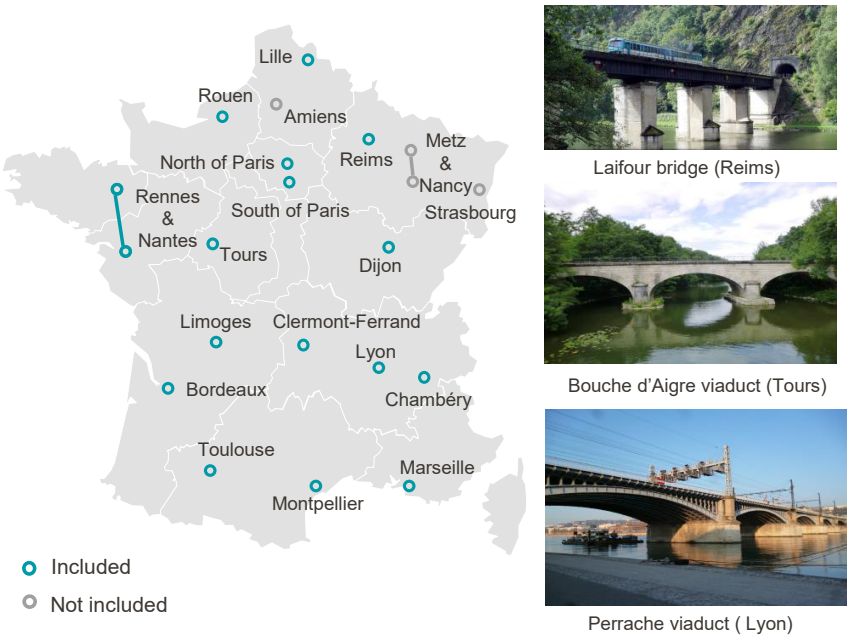


Figure 3.3: 19 SNCF regional offices and examples of bridges included in the dataset.

Figure 3.3 also illustrates examples of bridges included in the dataset. As mentioned in Chapter 2, bridges in TER or Intecités lines were mostly constructed 120 years ago. Due to the historic reasons, the construction techniques for these bridges are considered rather unified. Most bridges were built on timber piles, caissons, concrete or cement foundations. Lots of them are masonry arch bridges. Several are constructed in concrete or steel.

3.2.1 Input data

Factors directly or indirectly related to scour are input data. To collect input data, we firstly refer to the inspection reports written by the engineers or inspectors. In these documents, general information regarding the asset, history, and observed damages around the rail asset are noted. Besides, some open-source data is included as well to complete the missing information in the reports.

3.2.2 Output data

Scour risk is the output of the machine learning model. After field inspections, the status of the asset (piers, abutments, or walls) should be evaluated. The assets are graded from 0 (totally damaged) to 10 (intact). Guidance is proposed to unify the standard of grading (see Appendix A). The proposed grade is then related to scour risk. Table 3.1 shows the scour risk classification rule.

Table 3.1: Scour risk classification rule

Class	**Grade**
High scour risk	$0 \leq \text{grade} \leq 5$
Low scour risk	$5 < \text{grade} \leq 10$

After finishing data collection, Table 3.2 shows the size of two datasets and the number of collected samples in each class. It can be seen that the size of the bridge pier dataset (208) is much larger than the Abutment&Wall (124). This is because for one bridge, it could have multiple piers but only two abutments. 75 bridges are covered in the pier dataset while 53 bridges and 7 walls for Abutment&Wall. It can be seen that the two scour risk classes are not equally distributed in pier dataset. This issue is addressed in subsection 6.2.3, Chapter 6, where the data oversampling technique is applied.

Table 3.2: Information of the two datasets

	Class distribution			
Dataset	**Low**	**High**	**Sum**	**Comments**
Pier	57	151	208	75 bridges
Abutment&Wall	58	66	124	53 bridges and 7 retaining walls

Raw data is collected at this stage. The following steps should be feature selection and data cleaning.

3.3 Feature selection

A risk analysis usually includes two key components: hazard and vulnerability (Goerlandt et al., 2015; Todinov, 2006). Hazard is defined as the condition (flood event) which could possibly cause the undesirable events. Vulnerability refers to the susceptibility of the receptor (rail infrastructure) to the damaging effects of a hazardous event. Because of the stochastic nature of hydrologic events and also the uncertainty about the watercourse - foundation - surrounding environment interaction, the driving forces to vulnerability are still unclear now (Bento et al., 2020). Studies have shown that bridge failures could be triggered by scour hole, debris accumulation and scour history. The uncertainty about scour history is in conjunction with potentially unknown foundation depth or state, bridge characteristics and number of floods in the last five years, which could increase the vulnerability (Argyroudis and Mitoulis, 2021). Therefore, main factors that could affect the scour risk were firstly divided into four categories, namely bridge characteristics, environment, history, and changing factors (see Figure 3.4).

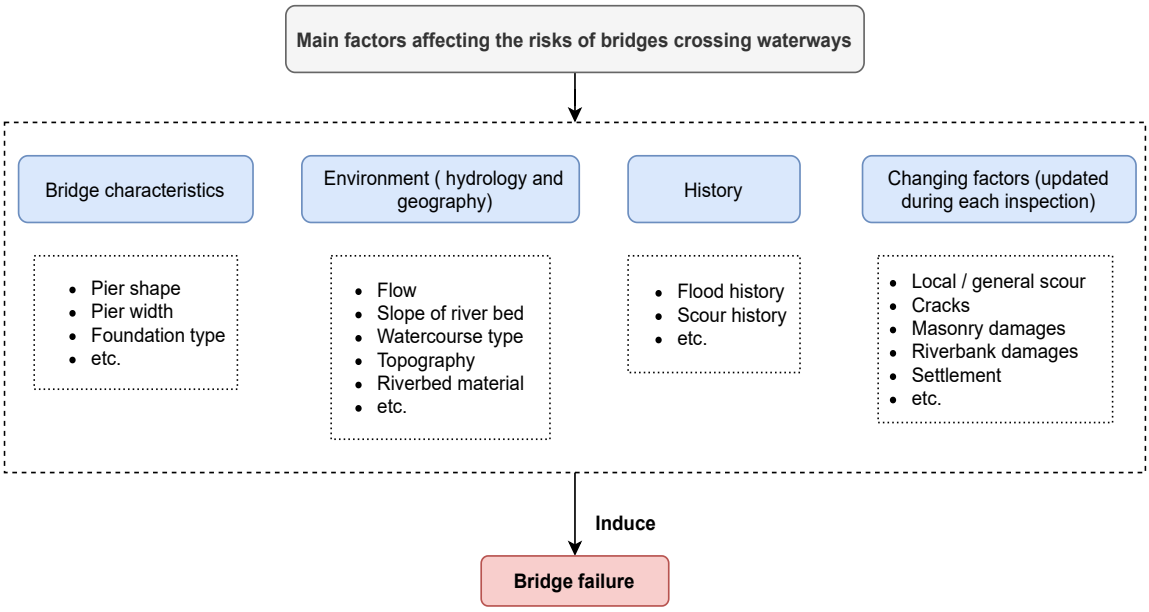


Figure 3.4: Main factors affecting the bridge failure due to scour.

It should be noted that only hazard and vulnerability factors for scour risk assessment are considered in this study. Consequence factors (e.g., the frequency of the train passed by each day, the difficulty to find an alternative transport when the traffic is disrupted) are not included since currently at SNCF, the railway engineers make decision only from a technical perspective. The consequence factors are considered afterwards while planning the maintenance work by the regional office.

In order to choose input parameters, current practical guidelines in France (ARPSA, Cerema, 2019; Livret A, SNCF, 2005), in Japan (scoring table, Takayanagi et al., 2018) and in the UK (EX2502, HR Wallingford, 1992) are referred to firstly. Brief presentations of ARPSA, scoring table and EX2502 are shown in section 2.4.

In the end, a dataframe consisting of 124 variables was proposed to cover as much information as possible. In this dataframe, essential and non-essential information for scour risk assessment is included, such as the region of the bridge, rail line number,

name of the structure, construction year, bathymetry data, and detailed description of observed damages. This dataframe cannot be used directly to train the machine learning model, considering the number of features we had and types of variables (both text and numerical variables are in it). Therefore, a feature selection work is necessary and is stared from this file by following both engineering and statistical perspectives.

Subsections 3.3.1 and 3.3.2 present the feature selection work for bridge pier. The Abutment&Wall dataset follows the same procedure and its feature selection process is presented briefly in subsection 3.3.3.

3.3.1 Engineering perspective

Compared with a field inspection which needs to note detailed information of the infrastructure and its surrounding environment, machine learning model performs a generic analysis and each variable needs to be representative enough for one kind of information. In this circumstance, before going to use statistical analysis for feature selection, the first step is to combine variables with similar information from an engineering perspective.

The original dataframe has flow depth and river width values measured from upstream, downstream, and under the bridge. For simplification, all these values are replaced as average flow depth and average river width.

Concerning the variables for damage descriptions, originally, there are 11 and 4 features for channel and riverbank respectively. Each variable is quite unique to describe one type of damage. For example, the 4 features describing the damages at riverbank are: excessive vegetation, debris, scour hole, landslide. These variables are then simplified as channel rating and river bank rating since in a ML model, the selected variable needs to be general and common enough. Instead of considering each

damage, a rating score is given in the end by counting the sum of damages. Similar actions are also done for variables describing damages of bridge foundation and its protection.

In the end, the original dataframe is simplified as follows. Although there is no rule to determine the relationship between the size of data and number of required variables, according to the examples that we've seen, the number of variables is still a lot. Therefore, a statistical analysis is conducted afterwards.

Selected variables after engineering analysis are:

- **Environment**: flow type, slope of riverbed (%), specific flood flow (m^3/s), width of valley/width of low flow channel, topography, hydraulic structure in the vicinity, sinuosity, riverbed material, river width (m), average flow depth (m), velocity (m/s), navigable, flow obstruction (%);
- **Bridge characteristics**: streambank protection, movement, pier shape, pier width (m), foundation type, existence of foundation scour countermeasures, watercourse countermeasures, attack angle;
- **History**: scour history, flood frequency;
- **Changing factors**: susceptible of scour, channel rating, bank rating, existence of dislocation or deformation (bin.), existence of dislocation or deformation (%), local scour (bin.), local scour (%), other damages.

Below are explanations of some features:

1. **Specific flood flow (m^3/s)** is calculated from the French project Relational System For Auditing The Hydromorphology of the Rivers (SYRAH-CE) (Valette and Cunillera, 2010). It represents the flow which didn't exceed 99% of the time on the curve of classified flows.

2. **Width of valley/width of low flow channel** (note as WV/WC hereafter) is the ratio between width of valley and width of low flow channel. Like specific flood flow (m^3/s), this hydromorphological value is obtained from the database SYRAH-CE. It describes the type of valley in which the river is located. A high WV/WC value represents a wide alluvial valley, which offers the possibility of the river to change the flow path. For the zone where the WV/WC is high, the risk of inundation could be high as well.

3. **Channel rating** notes the condition of river channel. The rating depends on the quantity of observed damages in the river channel. For example, the riverbed bathymetry revolution, presence of vortex, excessive vegetation, debris, etc.

4. Dislocation or deformation is the damage existing on masonry as shown in Figure 2.8 (a), or the protection of foundation. **Existence of dislocation or deformation (bin.)** is the binary form[1] of the damage and the possible choices are yes or no. **Existence of dislocation or deformation (%)** is calculated as the damaged surface/volume divided by the surface/volume of the masonry wall or gabion.

5. Local scour is normally the scour hole existing around the foundation, as shown in Figure 2.6 (c). **Local scour (bin.)** is the binary form of the damage and the possible choices are yes or no. **Local scour (%)** is the scoured surface/volume divided by the surface/volume of the foundation.

Bridge foundation depth, which is a key parameter in most of risk management procedures, is not included here because it is difficult to obtain this information in current SNCF database. The British Network has also reported the same issue for knowing the foundation depth (Dikanski et al., 2018). However, the observed damages in the field could, in a way, reflect the bearing capacity of the bridge foundation, which

[1]Binary form means that the variable only has two possible outcomes.

in the end represent the bridge foundation depth.

3.3.2 Statistical perspective

Despite the fact that there is no rule to determine the relationship between the number of variables and size of data, variables after engineering analysis are still quite a lot. The feature selection process is then conducted from a statistical perspective.

3.3.2.1 Variables with low variance or not enough records

Table 3.3 shows the features to be eliminated and the corresponding reasons. For the variable "hydraulic structure in the vicinity", it is used for knowing the type of hydraulic structures (e.g., weir, dam) near the bridge. In the end, only 9% of data records the corresponding values and others have the same values "No". In other words, only a little number of bridges have the hydraulic structure nearby and the majority don't have it. Thus, this variable has low variance. Since features with low variance don't meaningfully contribute to the model's predictive capability (Singh, 2021), that's the reason why this feature should be dropped.

Table 3.3: Eliminated features due to non sufficient values

Variable name	**Reason**
Hydraulic structure in the vicinity	Low variance
Attack angle	Low variance
Watercourse countermeasure	Low variance
Movement	Not enough records

Such issue is also observed for the features "attack angle " and "watercourse countermeasure". 9.1% of data has the attack angle different than 90 degree. Concerning "watercourse countermeasure", only 5% of data has the corresponding value for de-

scribing the countermeasure of riverbed (e.g., gabion), which means in this database, most watercourse is not protected by scour countermeasures.

The rest variables in Table 3.3 for describing the movement of bridge are also eliminated. Although they are important criteria for knowing the status of a bridge, in practice, only bridges susceptible of scour are implemented with sensors to monitor the movement of bridge pier, abutment and bridge deck. In our dataset, only 12.3% of data has the corresponding values for bridge movement.

3.3.2.2 Feature elimination based on statistical information

For numerical values, data points (e.g., values in rows) based on the distinct values in the given column is grouped. The statistical summary of the series for the generated groups can then be calculated (examples in Table 3.4 and Table 3.5). Data is grouped by based on the scour risk class (high or low scour risk). The % means how many of the values are less than the given percentile.

Table 3.4 shows the example when the variable can be eliminated based on the statistical information. It can be observed that there are not so much differences for low and high scour risk cases regarding the variable "flow depth (m)". On the contrary, Table 3.5 illustrates the example when the variable seems to be important at this stage for scour risk evaluation. It can be clearly observed that compared with low scour risk cases, high scour risk cases have larger values for specific flood flow (m^3/s) in terms of mean, 25%, 50%, and 75% values.

Table 3.4: Statistical information of the variable "flow depth (m)"

Class	**Mean**	**Min.**	**25%**	**50%**	**75%**	**Max.**
Low scour risk	3.27	0.17	1.07	2.5	4.97	10.00
High scour risk	3.18	0.15	1.05	2.5	4.97	10.29

Table 3.5: Statistical information of the variable "specific flood flow (m^3/s)"

Class	Mean	Min.	25%	50%	75%	Max.
Low scour risk	507.53	2.09	35.41	140.28	536.86	5457.26
High scour risk	1027.95	2.09	93.72	343.66	1235.20	5457.26

For categorical variables, which means the variable only has a limited and usually fixed number of possible values, it's difficult to calculate the statistical information. Figure 3.5 presents the example for the variable "navigable". The scour risk classes are firstly represented by the numerical number. 0 means low scour risk and 1 signifies high scour risk. Later, the average values with corresponding standard deviations of selected variable can be calculated. From Figure 3.5, it can be seen that the average scour risk score is almost the same for non-navigable and navigable groups. For this reason, this variable can be eliminated at this stage.

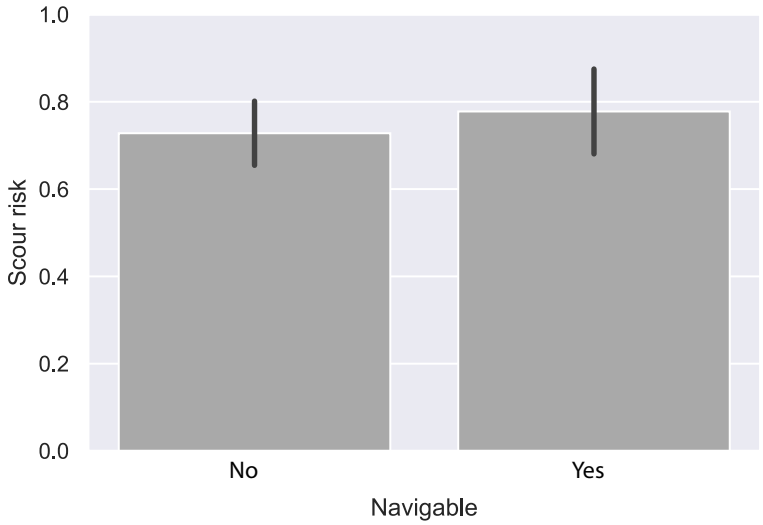


Figure 3.5: Bar chart for the variable "navigable".

Based on the above principles, eliminated features from the statistical perspective

are:

- **Environment**: river width (m), average flow depth (m), velocity (m), navigable, flow obstruction (%);
- **Bridge characteristics**: streambank bank protection, pier width (m);
- **Changing factors**: existence of dislocation or deformation (%), local scour (%).

To summarize, Table 3.6 shows the number of selected features after each step. A more detailed presentation of the 18 variables after statistical perspective selection is shown in section 3.5.

Table 3.6: Number of selected features after each step

Step	**Number of features**
Raw data	124
Engineering perspective	31
Statistical perspective (step 1)	27
Statistical perspective (step 2)	18

3.3.3 Feature selection for Abutment&Wall dataset

Abutment&Wall dataset has the same original database as the bridge pier. Since the number of data collected for the Abutment&Wall dataset is less than bridge pier dataset (208 for piers and 124 for Abutment&Wall dataset), the required features for training the ML model should be less as well, in order to avoid overfitting. In this circumstance, to start, the feature selection for Abutment&Wall dataset follows the same process as the bridge pier dataset.

Generally speaking, except the elimination of the variable "bridge pier shape",

other features dropped in Abutment&Wall dataset share the same reasons as the bridge pier dataset. The only difference exists in the feature "attack angle". It has a low variance in the bridge pier dataset. However, in the Abutment&Wall dataset, 44.7% of the data has the differentiated value. Thus, the feature "attack angle" is included in the selected features this time.

For both bridge pier and Abutment&Wall datasets, feature selection stops here. The remained features are used for training the machine learning model.

However, for Abutment&Wall dataset, the size of data is two times less than the bridge pier while the number of selected features is almost the same. It is very likely to have an overfitting model if training with current selected features. How to select the optimal features for the Abutment&Wall dataset will be discuss detailly in Chapter 4.

It should be noted that this feature selection process is based on the current dataset, which comprises only a limited number of bridges. More variables could possibly be included once more data is collected in the future. However, we try to make the current dataset representative enough by covering different geographical, hydrological backgrounds and structural types.

The final selected features used as input parameters and their detailed information are shown in section 3.5.

3.4 Data preprocessing

Once we've decided the input variables, data should be prepossessed. This section introduces the approaches to make the data ready to be used directly in the ML model.

3.4.1 Types of variables

The first step is to decide the type of variable. Among the selected features, only slope of riverbed, specific flood flow, WV/WC are numerical (Num.) variables and the rest are categorical ones. The categorical variables are also divided into three types: ordinal, nominal and binary.

Ordinal (Ord.) indicates that the variable has a clear ordering of the subcategories (for example, fluvial may induce less scour risk compared to torrential). Nominal (Nom.) variable, on the other hand, describes a variable without ordering or ranking subcategories (for example, different types of foundations are considered to have the same probability to induce scour risk). Binary (Bin.) is designated for a variable only having two subcategories. Figure 3.6 and Figure 3.8 list the types of each selected feature. The proposed order of subcategories were discussed with SNCF experts and may refer to the work of Deng and Cai (2010) and Wang et al. (2017).

3.4.2 Categorical data encoding

The ML model requires all input and output variables to be numeric. In other words, all categories must be converted to numeric form. An ordinal encoding involves mapping each category to an integer, and normally the integer value starts from zero.

For nominal variables there is no inherent relationship between categories. In case that the ML model will not confuse with an ordinal variable, the common solution is to use one-hot encoding, which means to create one binary attribute per category. For example, there are three possible values for the variable foundation type (B9 or BA9). A caisson mode foundation can then be interpreted into [1, 0, 0] ([0, 1, 0] and [0, 0, 1] are respectively for timber piles and mass concrete foundations). The encoding values of categorical variables are shown in Figure 3.6 and Figure 3.8.

3.4.3 Data scaling

Feature scaling is another essential step in data preprocessing. Despite the fact that data is not necessarily required to be in the same scale for tree-based algorithms (e.g., extreme gradient boosting, random forest) (Xia et al., 2017), to compare with other algorithms, 0-1 scaling is employed in our study.

Supposing that F=$\{X, Y\}$ denotes the whole dataset, where $X = (x_1, x_2, \ldots x_m)$ is the m-dimension input feature space and $Y = \{0, 1\}$ represents the binary output variables, namely low scour risk ("0") and high scour risk ("1") respectively. Input data is then computed by the equation below:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

where x' will be the novel input data after normalization. Equation 3.1 makes sure that all input data ranges from 0 to 1.

3.5 Final data for ML model

After feature selection in section 3.3 and data preprocessing in section 3.4, data is now ready to be used to build the ML model. But before training, this section presents the summary of the selected features as input variables and their relative information.

3.5.1 Bridge pier dataset

Input variables selected for bridge pier dataset are shown in Figure 3.6. For convenience, each variable is represented by its symbol hereafter. As mentioned in subsection 3.4.2, a ML model requires data to be numerical. The "Encoding" column in Figure 3.6 shows the numerical number of the values (categories) for categorical variables.

Group	Symbol	Variable	Type	Values/Range	Encoding
Environment	C1	Flow type	Ord.	Fluvial - Other - Torrential	0-1-2
	C2	Slope of riverbed (%)	Num.	[0.01-3.08]	-
	C3	Specific flood flow (m^3/s)	Num.	[2.09 - 5457.26]	-
	C4	Width of valley/Width of low flow channel	Num.	[1.52 - 226.68]	-
	C5	Topography	Ord.	Plain - Other - Mountain	0-1-2
	C6	Flow sinuosity	Ord.	Almost straight - Sinuous - Extremely sinuous	0-1-2
	C7	Riverbed material	Ord.	Rock - Cohesive soil - Cohesionless soil	0-1-2
Bridge characteristics	B8	Pier shape	Ord.	Triangular-nosed - Circular or oblong - Rectangular	0-1-2
	B9	Foundation type	Nom.	Caisson - Timber piles - Mass concrete	-
	B10	Existence of foundation scour countermeasures	Bin.	No - Yes	0-1
History	H11	Scour history	Bin.	No - Yes	0-1
	H12	Flood frequency	Ord.	Occasional - Frequent - Very often	1-2-3
Changing factors	I13	Susceptible of scour	Bin.	No - Yes	0-1
	I14	Channel rating	Ord.	Very good - Good - Fair - Poor - Very Bad	0-1-2-3-4
	I15	Riverbank rating	Ord.	Very good - Good - Fair - Poor- Very Bad	0-1-2-3-4
	I16	Existence of dislocation or deformation around masonry or gabion	Bin.	No - Yes	0-1
	I17	Existence of local scour	Bin.	No - Yes	0-1
	I18	Rating of other damages of foundation (corrosion, timber piles degradation, cracks, etc.)	Ord.	Very good - Good - Fair Poor	0-1-2-3

Figure 3.6: Data description - bridge pier dataset.

To see the distribution of data, the histograms of input parameters are shown in Figure 3.7.

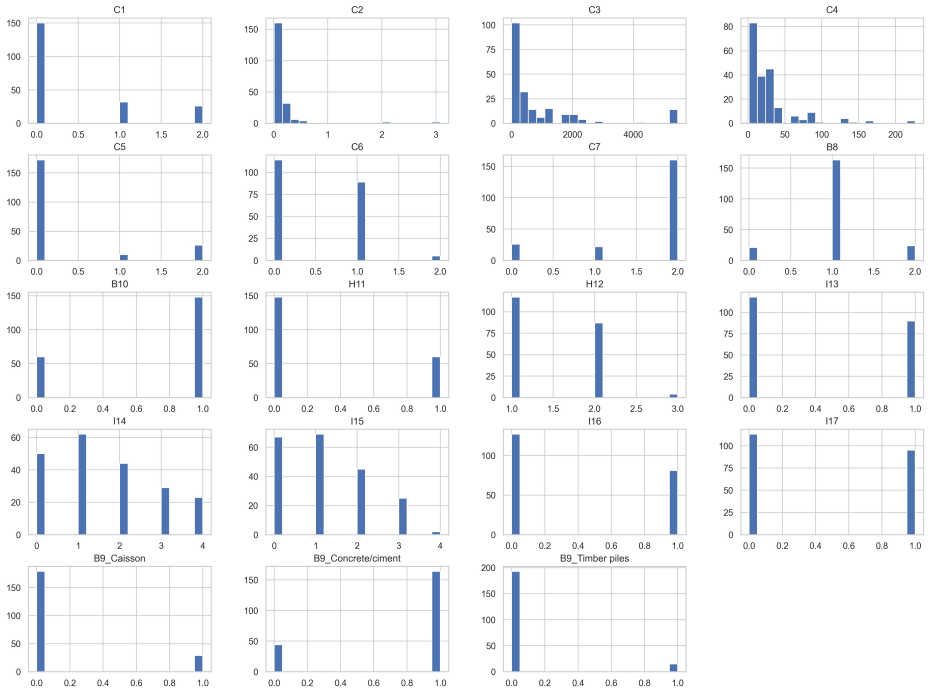


Figure 3.7: Histograms of input parameters - bridge pier dataset.

3.5.2 Abutment&Wall dataset

Figure 3.8 presents the selected features as input parameters and their relative information. The distribution of input parameters is shown in Figure 3.9.

Group	Symbol	Variable	Type	Values/Range	Encoding
Environment	CA1	Flow type	Ord.	Fluvial - Other - Torrential	0-1-2
	CA2	Slope of riverbed (%)	Num.	[0.01 - 11.72]	-
	CA3	Specific flood flow (m^3/s)	Num.	[0.39 – 5457.26]	-
	CA4	Width of valley/Width of low flow channel	Num.	[1.52 - 226.68]	-
	CA5	Topography	Ord.	Plain - Other - Mountain	0-1-2
	CA6	Flow sinuosity	Ord.	Almost straight - Sinuous - Extremely sinuous	0-1-2
	CA7	Riverbed material	Ord.	Rock - Cohesive soil - Cohesionless soil	0-1-2
Bridge characteristics	BA8	Foundation type	Nom.	Caisson - Timber piles - Mass concrete	-
	BA9	Existence of foundation scour countermeasures	Bin.	No - Yes	0-1
	BA10	Attack angle	Num.	[0.0 - 52.00]	-
History	HA11	Scour history	Bin.	No - Yes	0-1
	HA12	Flood frequency	Ord.	Occasional - Frequent - Very often	1-2-3
Changing factors	IA13	Susceptible of scour	Bin.	No - Yes	0-1
	IA14	Channel rating	Ord.	Very good - Good - Fair - Poor - Very Bad	0-1-2-3-4
	IA15	Riverbank rating	Ord.	Very good - Good - Fair - Poor- Very Bad	0-1-2-3-4
	IA16	Existence of dislocation or deformation around masonry or gabion	Bin.	No - Yes	0-1
	IA17	Existence of local scour	Bin.	No - Yes	0-1
	IA18	Rating of other damages of foundation (corrosion, timber piles degradation, cracks, etc.)	Ord.	Very good - Good - Poor	0-1-2

Figure 3.8: Data description - Abutment&Wall dataset.

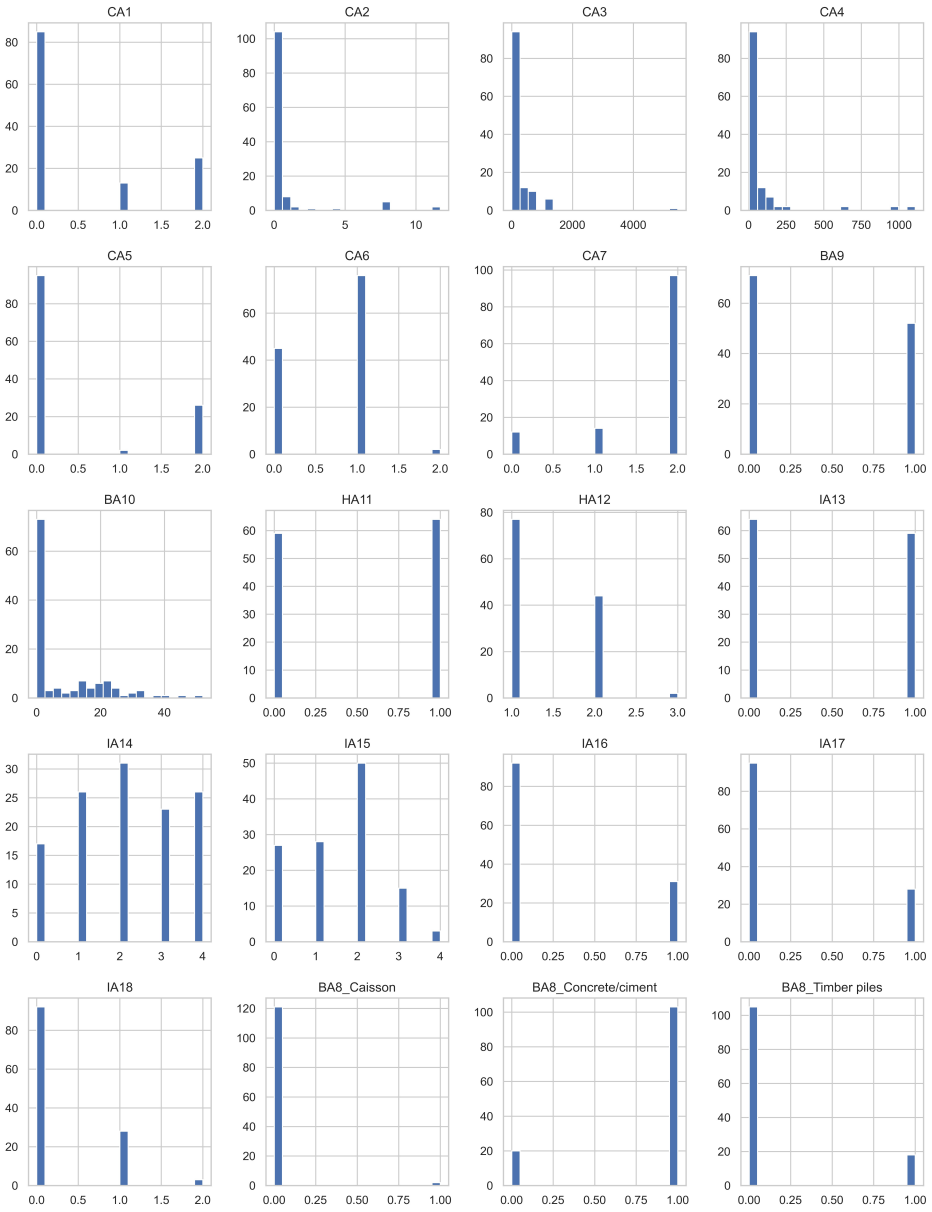


Figure 3.9: Histogram of input parameters - Abutment&Wall dataset.

3.6 Conclusions

In this chapter, the process to choose input parameters, and the approaches to make selected variables adaptable for ML models are presented. Two datasets are established in the end. One for bridge piers and another for bridge abutments, wing walls and retaining walls. It should be noted that the number of data in each dataset is not the same: 208 measurements for bridge pier dataset while 124 for the Abutment&Wall dataset.

Selected features for bridge pier and Abutment&Wall datasets are almost the same and they could all influence the scour risk. But it's noteworthy that features eliminated from engineering and statistical perspectives do not mean they don't have a great impact on scour. Since the objective of this work is to perform a ML-based analysis, variables who don't correspond to the ML feature selection policies (e.g., low variance, not enough records) must be dropped.

Furthermore, the current database has a relative small size and is constructed in a limited given time. Once more data is collected, it's possible to include more features.

Concerning the Abutment&Wall dataset, since its size is approximately two times smaller than pier dataset but the number of input variables are almost the same, an overfitting of model could possibly happen and it could also be the case for bridge pier dataset. But without training and test results, it is only the hypothesis since overfitting depends on the training and test results of the ML model, as presented in Chapter 2. Thus, we are going to build the ML model and the results will be presented in the next Chapter.

Chapter 4

Model construction

4.1 Introduction

In the last decade, machine learning was in a rapid development and seemed to be an alternative approach to overcome the limitations of empirical-based methods. Machine learning applications in the field of civil engineering are presented in section 2.6 and promising results are obtained.

In Chapter 3 section 3.3, we've defined the risk consists of two key components, namely hazard and vulnerability. Generally, the risk level is determined via the risk matrix (see Figure 4.1 (a)) and a threshold is then set to categorize the risk level. This procedure has also been followed in ARPSA (Cerema, 2019).

However, the category of hazard and vulnerability, as well as the threshold in risk matrix are primarily based on empiricism. Bridge scour induced by flood events is a rather complicated physical process. It is about the interaction among foundation, watercourse, and subsoil. Considering the stochastic nature of hydrologic events and the uncertainty about the structural status of the historical rail infrastructure, a

ML based model for risk prediction may be another option since it learns from the observed patterns without using predefined thresholds.

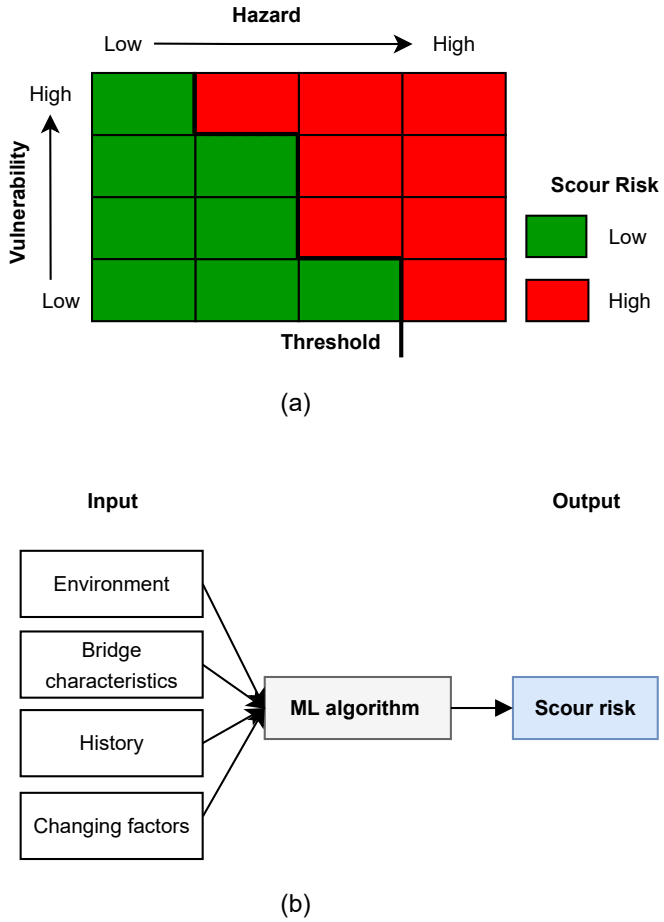


Figure 4.1: Risk evaluation through a risk matrix (a) and a ML model (b)

Figure 4.1 (b) shows the scour risk evaluation via a ML model. Considering the types of variables and the size of data we have, supervised learning is adopted in this study. The selected features after Chapter 3 are used as input parameters to train the model and the model output is the scour risk.

In this chapter, two machine learning models are constructed: one for bridge pier

dataset, and another for Abutment&Wall dataset. Section 4.2 presents briefly some popular and commonly used algorithms. Later, the training and test results of two datasets are shown in section 4.3 and section 4.4 respectively. All ML models' performance is examined through measurements introduced in section 2.5.3.3.

4.2 Brief introduction of applied machine learning algorithms

When looking at the literature, different machine learning models have been employed in the domain of civil engineering, like support vector machine (Hong et al., 2012), artificial neural network (Pala et al., 2008), recurrent neural network (Ninić et al., 2017), convolutional neural network (Zhao et al., 2020; Ran et al., 2019).

Besides, compared with single sophisticated learning algorithms such as decision trees or k-nearest neighbours, ensemble learning combines predictions from two or more models. Studies have shown that the integrated ensemble models achieved better results compared with a single ML model (Lessmann et al., 2015; Nanni and Lumini, 2009).

Extreme gradient boosting (XGBoost) algorithm is considered as one of the advanced supervised ensemble learning models. Proposed by Chen and Guestrin (2016), it was employed in various Kaggle machine learning competitions. In the research community, XGBoost algorithm was used to predict the undrained shear strength of soil (Zhang et al., 2021b), the concrete electrical resistivity (Dong et al., 2020), and the earth dam slope stability (Wang et al., 2020b), etc. In general, XGBoost model manifests an encouraging prediction capacity in diverse engineering problems.

In this circumstance, this section presents the basic principle of XGBoost algorithm. Other popular machine learning algorithms are also introduced to be compared with

the XGBoost classifier in the end.

4.2.1 Extreme gradient boosting

Extreme gradient boosting (XGBoost) is proposed by Chen and Guestrin (2016) based on tree models. It has been recognized in a great number of machine learning and data mining competitions. During the 2015 Kaggle challenge, nearly 60% published winning solutions used XGBoost.

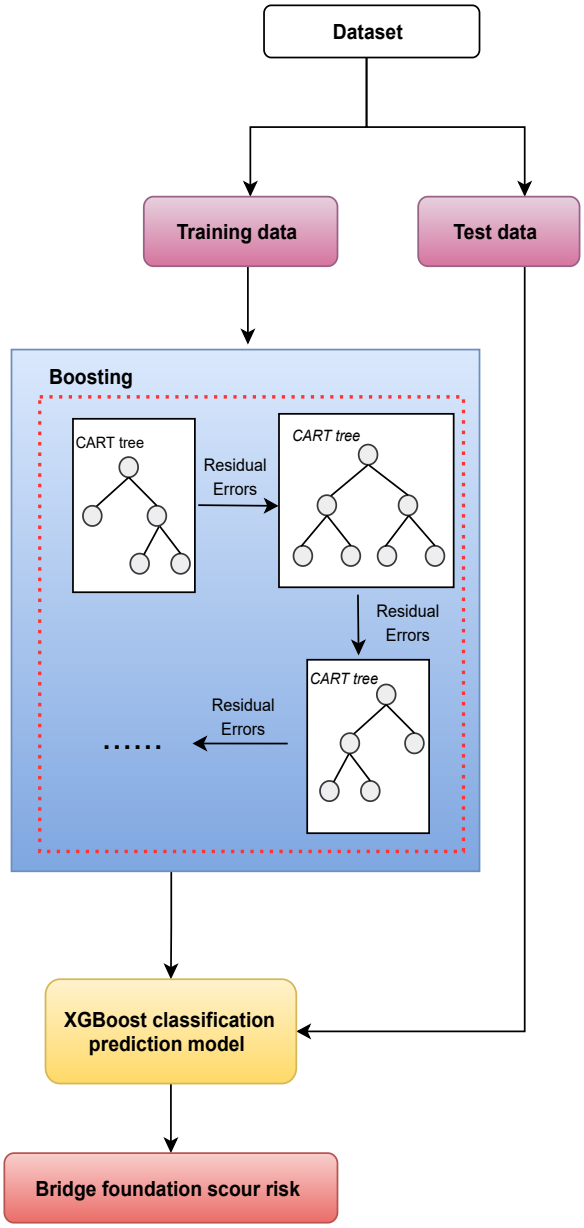


Figure 4.2: Schematic presentation of the XGBoost classifier.

Figure 4.2 depicts schematically the process of XGBoost algorithm. It is developed using gradient boosting algorithm under the ensemble learning framework. Compared with a single decision tree (also known as CART, Breiman et al., 1984), ensemble learning generates multiple decision trees to achieve high accuracy in prediction. In

gradient boosting algorithm, it combines several weak learners into a strong learner by sequentially adding predictors to correct its predecessor. The residual error is minimized by adding CART trees. Compared with traditional gradient boosting algorithm, XGBoost algorithm adds a regularized term to penalize the complexity of the model in order to avoid overfitting. A more detailed mathematical deduction concerning XGBoost algorithm may refer to Appendix B.

4.2.2 Support vector machine

Support vector machine (SVM) (Cortes and Vapnik, 1995) is a powerful supervised learning algorithm which is capable of classification and regression. The model is simply a linear function plus a bias term. To separate the data points, an optimal hyperplane is defined in SVM. The objective is to make sure that the hyperplane has the maximum distance between data points from both classes. Compared with other ML algorithms such as logistic regression, SVM addresses the model overfitting problem by balancing its complexity against its success at fitting the training data. In other words, the training data is allowed on the wrong side of the hyperplane (see Figure 4.3).

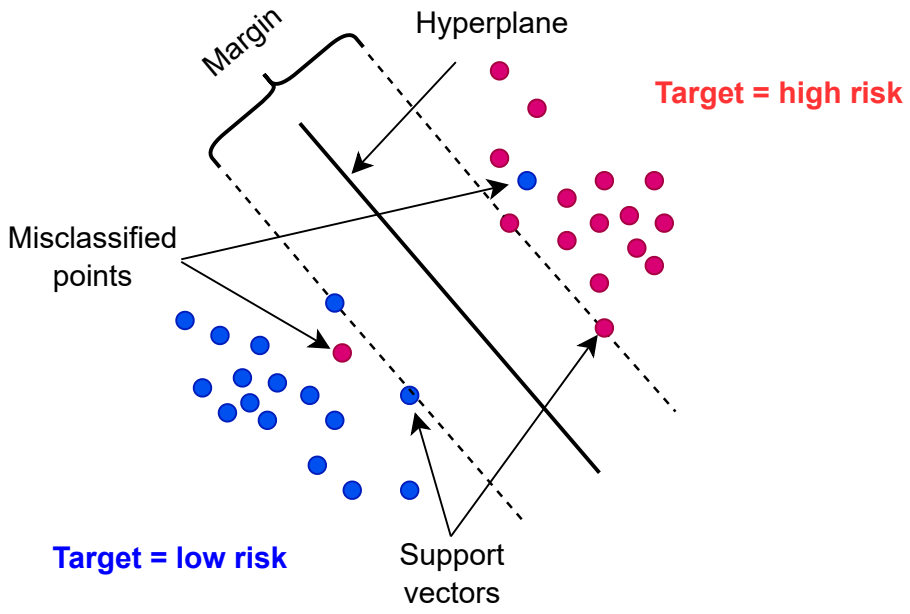


Figure 4.3: Schematic presentation of the SVM classifier

In order to overcome the limitations of a linear function, kernel trick is employed in SVM to implicitly map the training data into a higher dimension space but without complicating the computational process (Kordjazi et al., 2014). Numerous kernel functions were proposed in the literature (Cristianini and Shawe-Taylor, 2000), even though polynomial and radial basis function (RBF) are the two well exploited kernels for geotechnical engineering problems (Goh and Goh, 2007; Samui, 2008; Samui et al., 2008). The present work used RBF kernel in the SVM model.

4.2.3 Random forest

Developed by Breiman (2001), random forest (RF) is another ensemble learning method that aggregates numerous decision trees. RF features advantages for classification and regression problems in many aspects. A single tree is a weak classifier due to its high variance. To overcome this shortcoming, RF forms a forest by generating

a great amount of decision trees. The objective of RF is to grow each decision tree individually by adopting the re-sampling method called bagging. Bagging makes each tree uncorrelated in order to avoid overfitting. By generating lots of trees, a more robust model could be achieved by obtaining low variance as well as the increase of model accuracy (Breiman, 2001). The schematic presentation of RF algorithm is shown in Figure 4.4.

The decision on class affiliation ("high scour risk"-"low scour risk") is counted by the majority vote among all trees. To construct a robust RF model, two priori hyper-parameters should be optimized: the number of trees in the forest and the minimum number of number of samples required to split an internal node (Rodriguez-Galiano et al., 2012). These two hyper-parameters help to minimize the error as well as to obtain a satisfactory model performance.

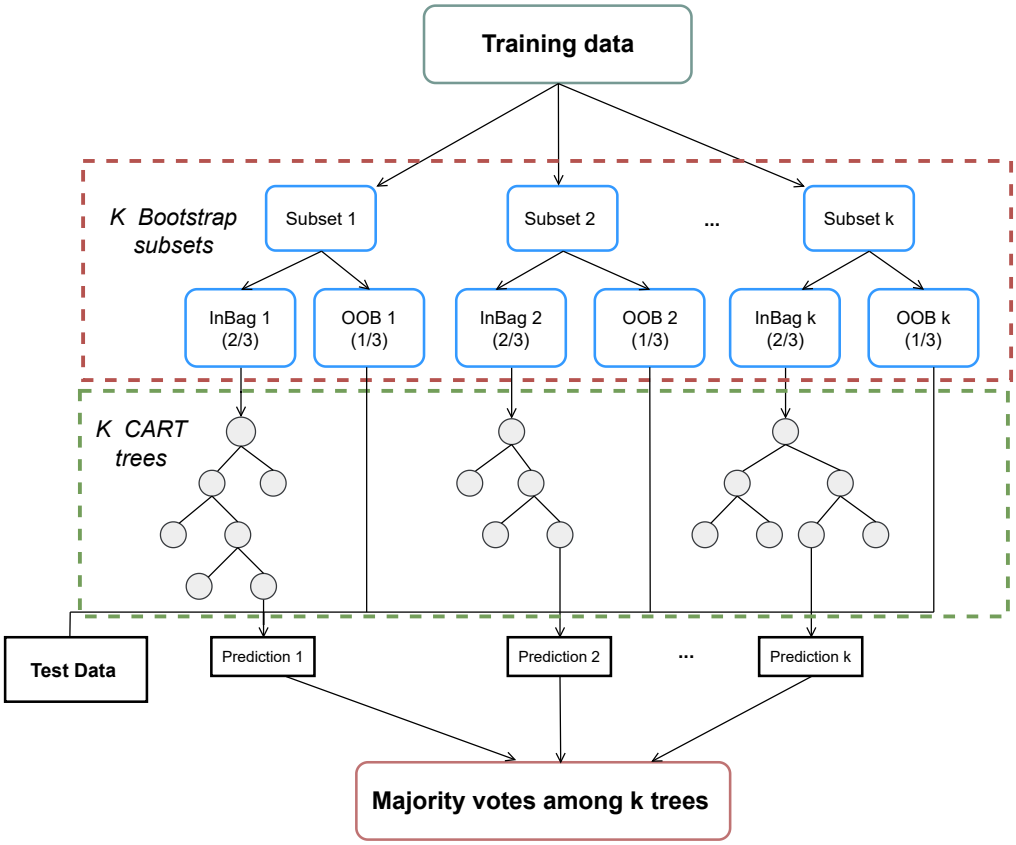


Figure 4.4: Schematic presentation of the RF algorithm.

4.2.4 Multilayer perceptron

The multilayer perceptron (MLP) is perhaps the most popular and commonly used artificial neural network (ANN) structure (Liu et al., 2021). An MLP should include no less than three layers: an input layer, an output layer and one or more hidden layers. As illustrated in Figure 4.5, the information will pass from input layer to output layer through hidden layer(s). Therefore MLP is a feedforward neural network (FNN). Except the nodes in the input layer, each node multiplies every node in the previous layer by its interconnection (synaptic) weights and then adds the sum of the product. Later the sum passes through a nonlinear activation function (Zounemat-

Kermani et al., 2009). The obtained results after the activation function then pass on to the next layer as a new "input" and the aforementioned process will be repeated (Franklin, 2010; Liu et al., 2021; Olden and Jackson, 2002). The value of the output node is the weighted sum of previous hidden nodes. To minimize the cost function in MLP, Rumelhart et al. (1986) proposed the backpropagation (BP) algorithm. BP propagates the error from output node to the input nodes. It calculates the partial derivative of the cost function with respect to any weight and bias in the neural network. Moreover, the optimal value for weight and bias in the network are computed by gradient descent algorithm.

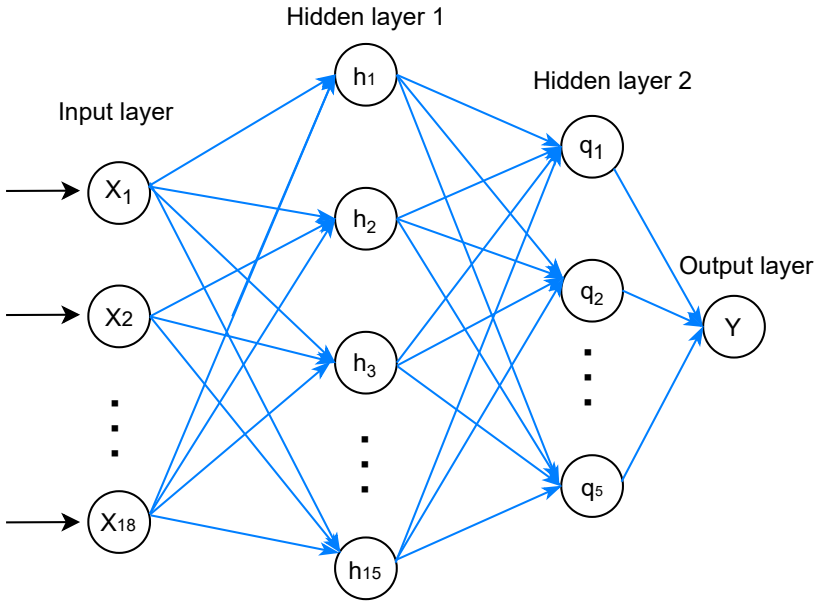


Figure 4.5: Architecture of MLP in the present study.

In this study, an MLP neural network with two hidden layers was established. There were respectively fifteen and five nodes in each hidden layer. The output node is the foundation scour risk. 20% dropout rate (Srivastava et al., 2014) was applied in the training phase in the aim of avoiding overfitting.

4.3 Bridge pier dataset

This section introduces the process to build the ML model for the bridge pier dataset. The research approach is presented schematically in Figure 4.6 and is described as follows:

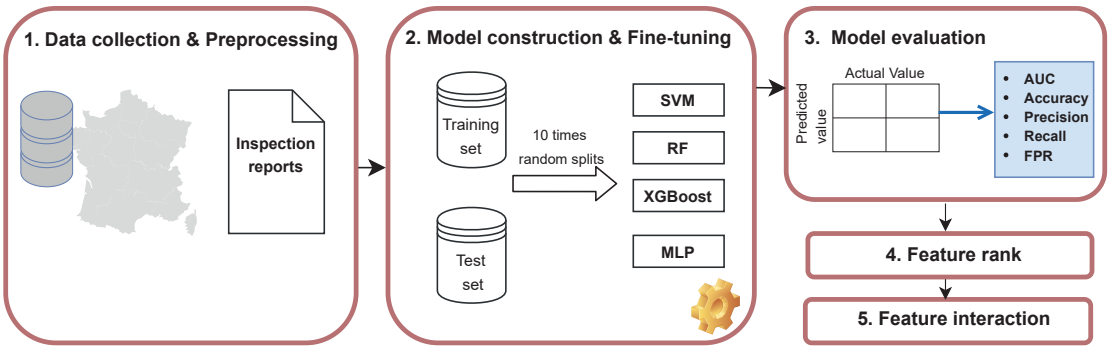


Figure 4.6: Schematic presentation of bridge pier dataset training process.

1. Data collected from SNCF inspection reports should be pre-processed to make sure the format was suitable for ML models. This part has already been presented in Chapter 3.

2. Later, 70% of the data was used to train the model and the remaining 30% was used for evaluation. Four popular classification algorithms presented in section 4.2 are applied and they respectively are: support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost) and multilayer perceptron (MLP). In order to see the robustness of each model, the random training-test data splits were repeated 10 times.

3. Then, ML models were evaluated by a confusion matrix introduced in section 2.5.3.3 between the predicted value and actual value. Five model performance measurements were then calculated, and they, respectively, are: area under the ROC curve (AUC), accuracy, precision, recall, and false positive rate (FPR).

4. Once the model was trained, the XGBoost algorithm based model was used to plot feature importance rank.

5. Lastly, to quantify the contribution and interaction among features, the XGBoost classifier, due to its high accuracy score, was trained and tested using one single input parameter each time.

The following sections present the training and test results as well as the feature importance discussion.

Concerning the programming environment, all ML models were programmed on Python 3.8.5 and built using different python libraries. Data was preprocessed using NumPy (Harris et al., 2020) and pandas libraries (The pandas development team, 2020). The XGBoost algorithm was applied using XGBoost package (Chen and Guestrin, 2016); the SVM and RF were operated using Sklearn library (Pedregosa et al., 2011); and the MLP was performed on Keras library (Chollet et al., 2015). Grid search was adopted in each algorithm to search for the optimal hyper-parameters.

4.3.1 Training and test results

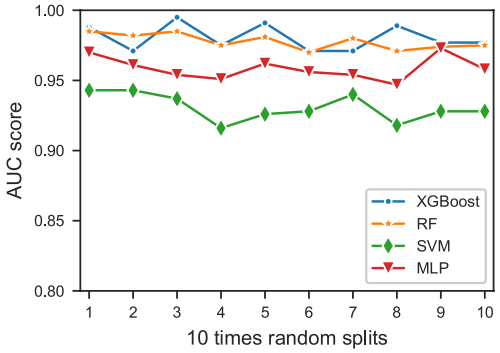
To build the model for prediction, four algorithms are applied and the random training-test splits are repeated 10 times. Figure 4.7 presents training and test scores of the model evaluation measurements in each split. These measurements are calculated from a confusion matrix, which is presented in section 2.5.3.3 by comparing the predicted class with actual class.

AUC measures how well the classifier distinguishes between the two classes. It is the area under the ROC curve (see Figure 2.19). The optimal value for AUC is equal to 1.

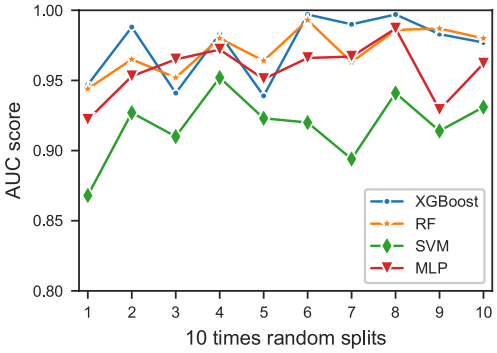
Accuracy describes how the model performs across the two classes. Precision mea-

sures the model's accuracy in classifying a sample as positive (high scour risk) class. Recall measures the model's ability to detect positive samples (high scour risk). A high recall classifier means that it can detect all high scour risk bridges, but at the same time, low scour risk examples are evaluated as high. The optimal values for accuracy, precision, recall are equal to 1.

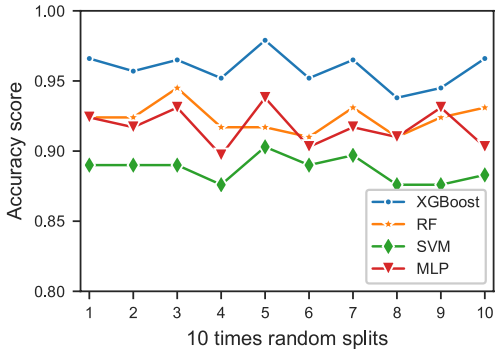
False positive rate (FPR) is the probability of false alarm. It measures among the predicted high scour cases: how many are predicted incorrectly (actually at low risk). The optimum value for FPR is 0.



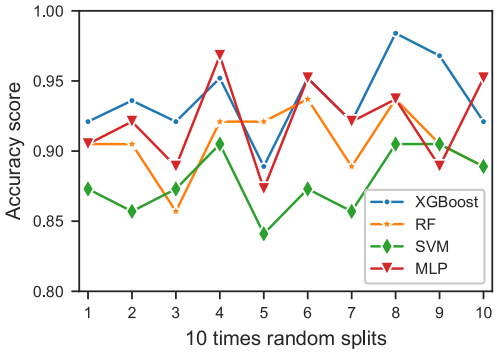
(a) AUC training scores



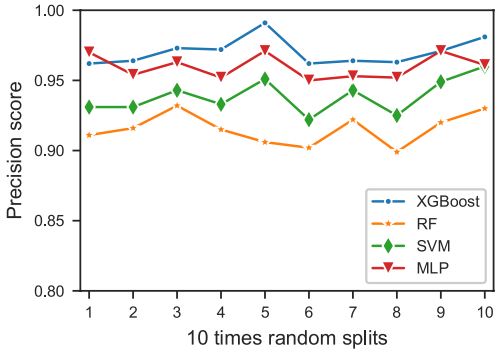
(b) AUC test scores



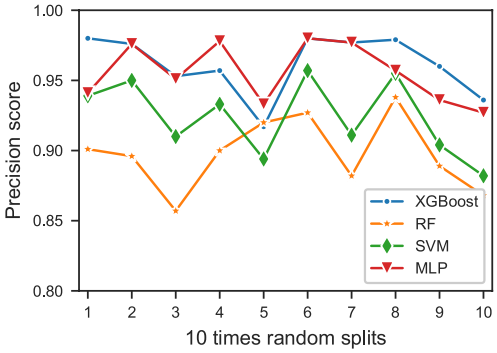
(c) Accuracy training scores



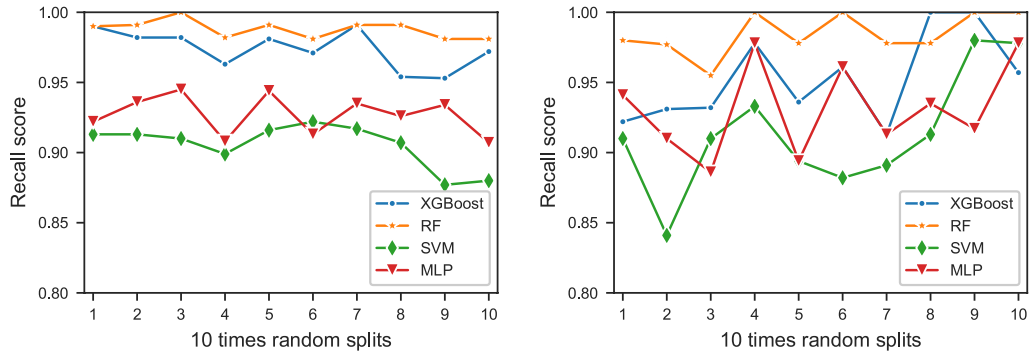
(d) Accuracy test scores



(e) Precision training scores

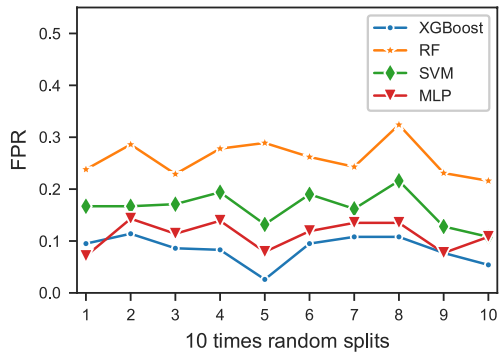


(f) Precision test scores

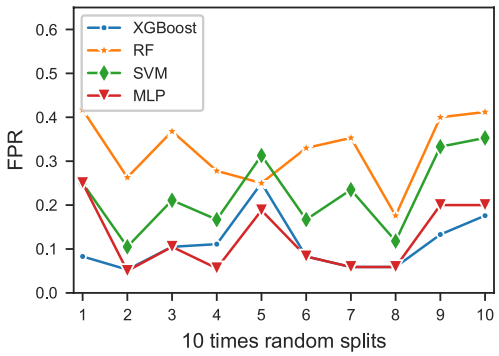


(g) Recall training scores

(h) Recall test scores



(i) FPR training scores



(j) FPR test scores

Figure 4.7: Training and test scores under 10 times random splits (bridge pier dataset)

From Figure 4.7 (a) - (f), it is observed that XGBoost has almost the highest AUC, accuracy, precision scores over the training and test sets in each split. Figure 4.7 (g) - (j) indicates that although RF classifier has a high recall score, its FPR is also high among the four algorithms.

Later, the average and standard deviation (STD) values of model performance measurements for training and test data are summarized in Table 4.1 and Table 4.2 respectively.

Table 4.1: Results on training data over model performance measures (pier dataset)

	AUC	**Accuracy**	**Precision**	**Recall**	**FPR**
XGBoost	0.981±0.009	0.959±0.012	0.970±0.010	0.974±0.014	0.085±0.057
RF	0.978±0.006	0.923±0.011	0.915±0.006	0.988±0.034	0.260±0.006
SVM	0.931±0.010	0.887±0.009	0.939±0.012	0.905±0.015	0.164±0.033
MLP	0.959±0.008	0.917±0.014	0.960±0.009	0.927±0.014	0.112±0.028

Table 4.2: Results on test data over model performance measures (pier dataset)

	AUC	**Accuracy**	**Precision**	**Recall**	**FPR**
XGBoost	**0.974 ± 0.023**	**0.938 ± 0.023**	**0.961 ± 0.022**	0.956±0.031	0.114±0.065
RF	0.971±0.016	0.907±0.026	0.897±0.027	**0.985 ± 0.016**	0.314±0.078
SVM	0.918±0.024	0.878±0.024	0.922±0.028	0.914±0.045	0.222±0.093
MLP	0.957±0.020	0.922±0.028	0.957±0.021	0.930±0.035	**0.111 ± 0.066**

It can be seen that among all four ML models, XGBoost classifier achieves the highest average score in terms of AUC, accuracy and precision for test data. Moreover, it has approximately the same lowest FPR score as MLP. RF model has the highest recall score but it also has the highest FPR, which indicates that it will have a higher probability for false alarms. In practice, this overestimated classifier may relatively require more unnecessary maintenance work and increase the maintenance cost in the end.

To conclude, the training and test results show that ML techniques could be successfully used for bridge scour risk prediction over the bridge pier dataset. Among the four proposed algorithms, XGBoost classifier has a very satisfactory performance over the two classes. Although RF model can predict well the high scour risk cases, it has the highest false positive rate as well. More studies are required to compare

comprehensively these two models, and it will be presented in Chapter 5.

4.3.2 Feature importance and interaction

As mentioned earlier, the evaluation of bridge scour involves information coming from different aspects. Feature importance, in this case, could provide vital guidance in terms of model interpretability. A trained XGBoost-based model is capable of automatically providing estimates of feature importance. Figure 4.8 illustrates features F weight scores. They are calculated as the total number of times a particular feature to split the tree in XGBoost.

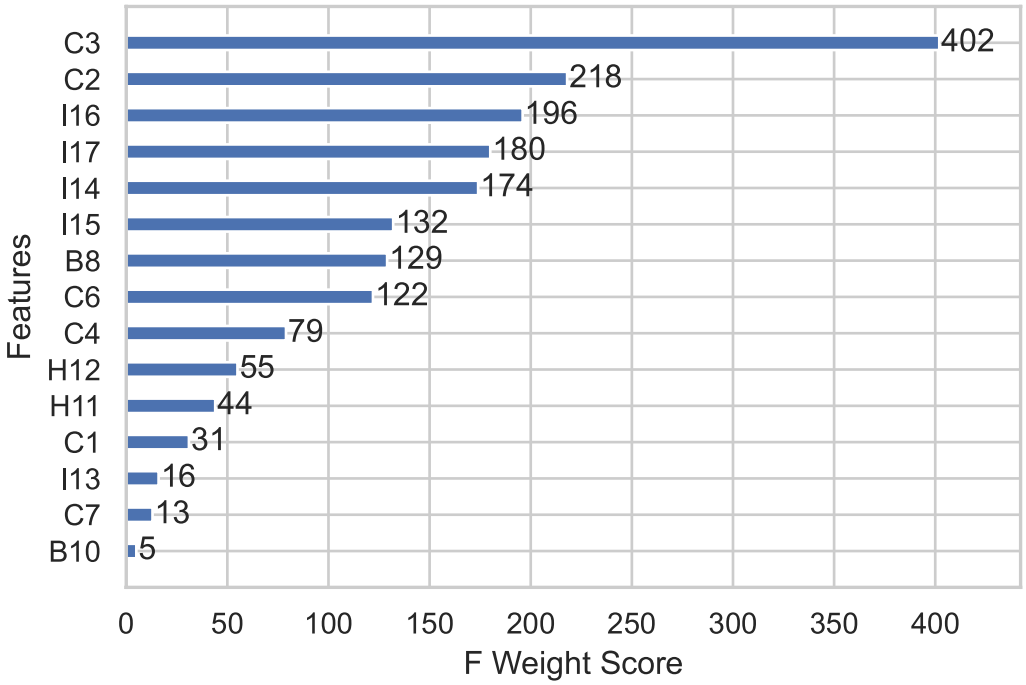


Figure 4.8: Feature importance plotted by the XGBoost algorithm (bridge pier dataset).

As shown in Figure 4.8, the feature C3 (specific flood flow (m^3/s)) is the most important variable which was considerably used (two times more used than second one C2) compared to the rest. It is then followed by features C2 (slope of riverbed

(%)), I16 (existence of dislocation or deformation around the masonry or gabion) and I17 (existence of local scour). Among the four most important variables, the top two C3 (specific flood flow (m^3/s)) and C2 (slope of riverbed (%)) belong to the category environment. In other words, besides the damages observed from each visual inspection, river hydrological characteristic and morphological regime (features C3 and C2) play a very important role in bridge scour risk evaluation. This result implies significant guidance for current industry process.

Besides, it should be also noted that two features C5 (topography) and B9 (foundation type) are not included in the ranking. By way of explanation these two features were not used at all as criteria to grow trees.

The choice of input features in this study was initially discussed with experts and followed the current practical guidance. C5 (topography) and B9 (foundation type) are two commonly presented features in lots of references. It is reasoned that C5 (topography) was not used in XGBoost model because the information was already included in C2 (slope of riverbed (%)), C4 (WV/WC) and C6 (flow sinuosity). As for B9 (foundation type), it could be due to that the engineer evaluates the scour risk in an asset-specific view. Consequently, knowing the foundation type of each bridge pier shall be indispensable whereas a ML model performs a generic and statistical analysis. Foundation type in this case could be seen as an unimportant factor when facing with a great number of cases.

A more detailed explanation regarding the feature importance discussion is presented in Chapter 6, where engineers' expertise is compared with the interpretation results from XAI.

Furthermore, to quantify the contribution of each feature to scour risk, an XGBoost model was built by using only one single feature each time. The AUC score for test data is shown in Figure 4.9.

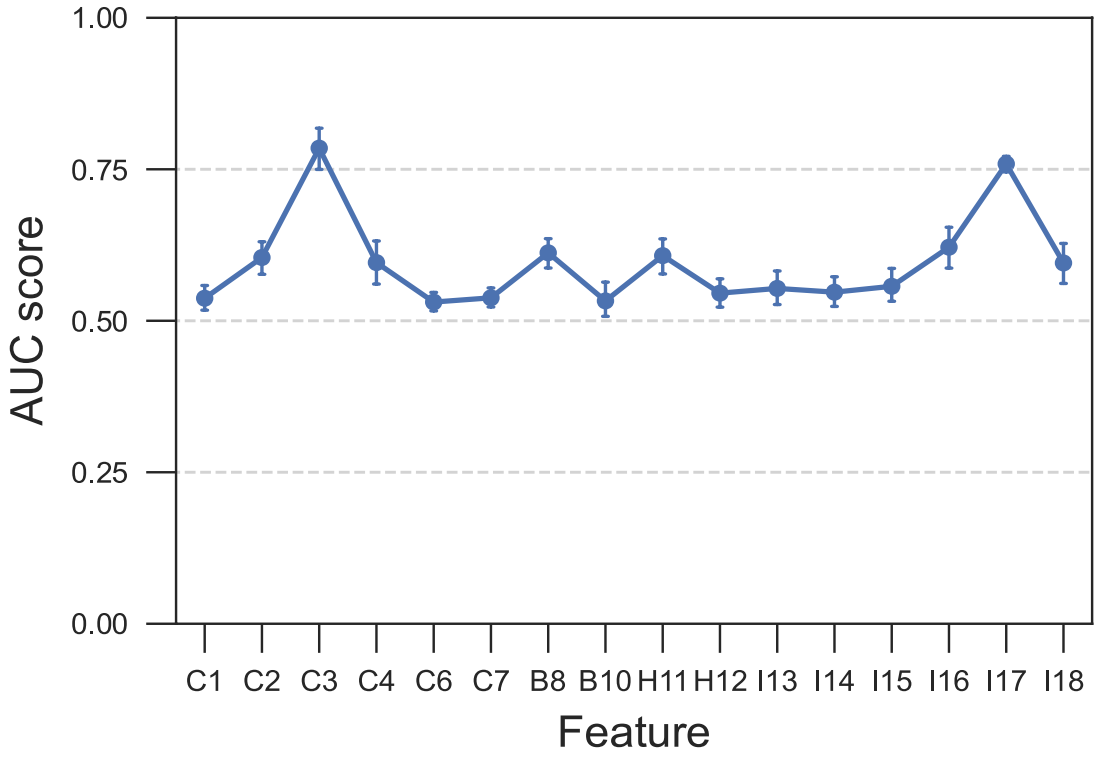


Figure 4.9: Single feature AUC score using XGBoost model

The two features C3 (specific flood flow (m^3/s)) and I17 (existence of local scour) have almost the same and the highest AUC score compared with others. However, it can be seen from Figure 10 that with one single feature, the classifier has a rather non-satisfactory performance even with the feature having a high rank. Furthermore, the AUC score is around 0.5 in most cases, which indicates that it works like a random classifier. The low AUC score shown in Figure 4.9 proves that one feature may not be adequate enough. It's the interaction among different features that reveals the mechanism of scour around bridge piers and makes sure the classifier has encouraging prediction results.

4.4 Abutment&Wall dataset

The Abutment&Wall dataset follows the same methodology as in bridge pier. Data after preprocessing is used to train the model. Considering the number of data we have (only half of the pier cases) and the performance of algorithms in pier dataset, RF and XGBoost are selected for training.

In this section, the feature selection work is presented at first in subsection 4.4.1. Later, the training and test results are shown in subsection 4.4.2.

4.4.1 Input parameters selection

The number of selected features for Abutment&Wall dataset after section 3.3 are the same as the pier dataset (18 in total) but the size of data is much less (208 for bridge pier but only 124 for Abutment&Wall). After feature selection in Chapter 3, we have doubts that overfitting could happen.

To begin, the model is trained with all selected features from Figure 3.8. Similar to the training process of the bridge pier dataset, 70% of the data was used to train the model and the rest for testing. Results are shown in Table 4.3 and Table 4.4.

Table 4.3: Abutment&Wall dataset training results with all selected features

	AUC	**Accuracy**	**Precision**	**Recall**	**FPR**
XGBoost	0.893	0.896	0.902	0.915	0.128
RF	0.907	0.907	0.889	0.930	0.117

Table 4.4: Abutment&Wall dataset test results with all selected features

	AUC	**Accuracy**	**Precision**	**Recall**	**FPR**
XGBoost	0.763	0.765	0.778	0.737	0.211
RF	0.736	0.742	0.810	0.739	0.267

Overfitting, which is presented in subsection 2.5.3.2, is defined as the machine learning model gives accurate predictions for training data but not test data. From Table 4.3 and Table 4.4, it can be clearly observed the huge gap between training and test results (overfitting) in both XGBoost and RF classifiers. Subsection 2.5.3.2 mentions that one of the techniques to avoid overfitting is to identify the most important parameters and train the model afterwards with simplified variables.

The feature importance of XGBoost classifier is plotted firstly and is shown in Figure 4.10. It is found that several variables are not mentioned in this rank (e.g., foundation type (BA8), topography (CA5), sinuosity (CA6)), which means they are not used for prediction. On the other hand, some variables in Figure 4.10 have relatively low feature importance scores (e.g., flow type (CA1), foundation scour countermeasure (BA9), flood frequency (HA12)). Thus, we decided to exclude variables that have little contribution to the prediction and see the performance of the two classifiers.

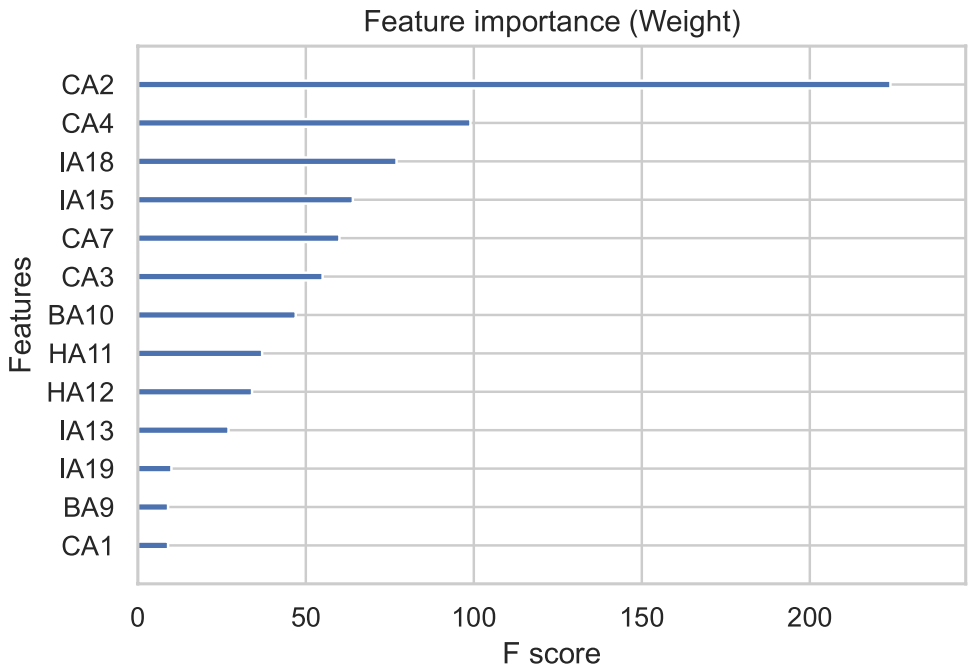


Figure 4.10: Feature importance given by the XGBoost algorithm (Abutment&Wall dataset).

Several tests were realized to determine the optimal variables for Abutment&Wall dataset. The excluded variables in each case are shown in Table 4.5. At the beginning, there is no excluded variable and all the input parameters shown in Figure 3.8 are used to train the model (Case No.1). From Case No.2, one variable is removed each time. Variables from Case No.2 to Case No.4 are those not shown in Figure 4.10. From Case No.5 to Case No.9, we exclude the variable each time by following the feature importance. It should be noted that variables in the category changing factors, namely rating of other damages of foundation (IA19) and susceptible of scour (IA13) are not excluded despite their relatively low feature importance score. We think the observations from the field are rather important.

Table 4.5: Cases for selecting the input variables regarding the Abutment&Wall dataset

Case No.	**Excluded variables**
1	None
2	BA8 (foundation type)
3	BA8 (foundation type), CA5 (topography)
4	BA8 (foundation type), CA5 (topography), CA6 (sinuosity)
5	BA8 (foundation type), CA5 (topography), CA6 (sinuosity), CA1 (flow type)
6	BA8 (foundation type), CA5 (topography), CA6 (sinuosity), CA1 (flow type), BA9 (foundation scour countermeasure)
7	BA8 (foundation type), CA5 (topography), CA6 (sinuosity), CA1 (flow type), BA9 (foundation scour countermeasure), HA12 (flood frequency)
8	BA8 (foundation type), CA5 (topography), CA6 (sinuosity),CA1 (flow type), BA9 (foundation scour countermeasure), HA12 (flood frequency), HA11 (scour history)
9	BA8 (foundation type), CA5 (topography), CA6 (sinuosity), CA1 (flow type), BA9 (foundation scour countermeasure), HA12 (flood frequency), HA11 (scour history), BA10 (attack angle)

By following these principles, Table 4.6 and Table 4.7 show the training and test results of each case by adopting the XGBoost and RF algorithms. The accuracy scores regarding each case are then illustrated in Figure 4.11.

Table 4.6: Training results of XGBoost and RF classifiers (Abutment&Wall dataset)

Case No.		**Accuracy**	**Precision**	**Recall**	**FPR**	**AUC**
1	RF	0.907	0.889	0.930	0.117	0.907
	XGBoost	0.896	0.902	0.915	0.128	0.893
2	RF	0.919	0.875	0.977	0.140	0.919
	XGBoost	0.837	0.842	0.750	0.150	0.871
3	RF	0.953	0.933	0.977	0.070	0.953
	XGBoost	0.872	0.889	0.870	0.125	0.872
4	RF	0.907	0.894	0.933	0.122	0.906
	XGBoost	0.895	0.909	0.889	0.098	0.896
5	RF	0.884	0.889	0.889	0.122	0.883
	XGBoost	0.919	0.923	0.900	0.065	0.917
6	RF	0.965	0.955	0.977	0.047	0.965
	XGBoost	0.907	0.947	0.857	0.045	0.906
7	RF	0.895	0.943	0.825	0.043	0.891
	XGBoost	0.942	0.949	0.925	0.052	0.941
8	RF	0.953	0.971	0.920	0.057	0.950
	XGBoost	0.930	0.925	0.945	0.065	0.930
9	RF	0.860	0.938	0.750	0.043	0.853
	XGBoost	0.917	0.913	0.925	0.065	0.930

Table 4.7: Test results of XGBoost and RF classifiers (Abutment&Wall dataset)

Case No.		**Accuracy**	**Precision**	**Recall**	**FPR**	**AUC**
1	RF	0.742	0.810	0.739	0.267	0.736
	XGBoost	0.765	0.778	0.737	0.211	0.763
2	RF	0.630	0.714	0.652	0.400	0.626
	XGBoost	0.684	0.682	0.750	0.389	0.681
3	RF	0.684	0.789	0.652	0.267	0.693
	XGBoost	0.684	0.682	0.750	0.389	0.681
4	RF	0.677	0.654	0.850	0.529	0.660
	XGBoost	0.676	0.682	0.750	0.412	0.669
5	RF	0.662	0.625	0.750	0.529	0.610
	XGBoost	0.811	0.909	0.800	0.167	0.817
6	RF	0.763	0.850	0.739	0.200	0.770
	XGBoost	0.894	0.954	0.875	0.071	0.902
7	RF	0.730	0.941	0.640	0.083	0.778
	XGBoost	0.838	0.913	0.840	0.167	0.837
8	RF	0.784	0.947	0.720	0.083	0.818
	XGBoost	0.811	0.909	0.800	0.167	0.817
9	RF	0.757	0.944	0.680	0.083	0.798
	XGBoost	0.817	0.916	0.815	0.185	0.762

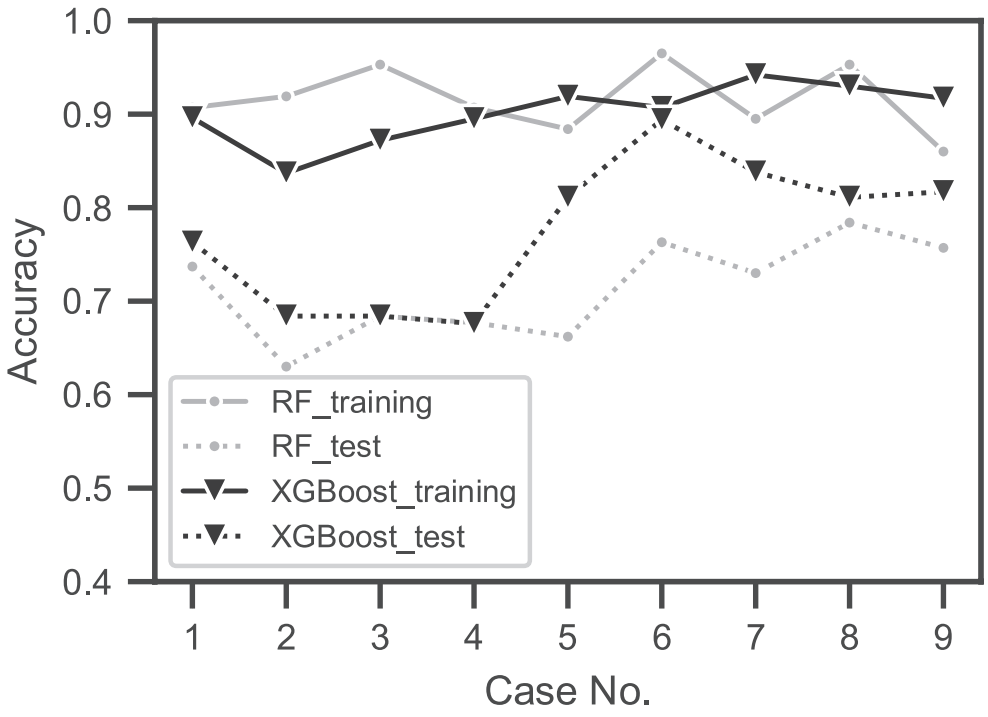


Figure 4.11: Accuracy score of RF and XGBoost classifiers.

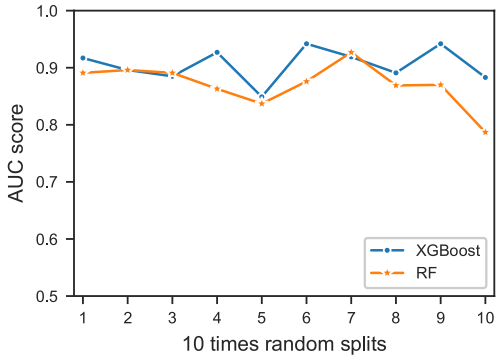
From Figure 4.11, it can be clearly observed that the accurate predictions over training but not test set from Case No.1 to No.5 in both XGBoost and RF algorithms. In Case No.6, XGBoost algorithm performs very well and has the least difference between training and test sets. From Case No.7 to Case No.9, the differences between the two sets become bigger but still, they are smaller than Case No.1 - No.5.

Figure 4.11 confirms the idea that overfitting could happen with initially selected features. Features in Case No.6 seem to be the optimal ones for the Abutment&Wall dataset. Based on these results, we are going to repeat the training and test process to see the performance of the two classifiers.

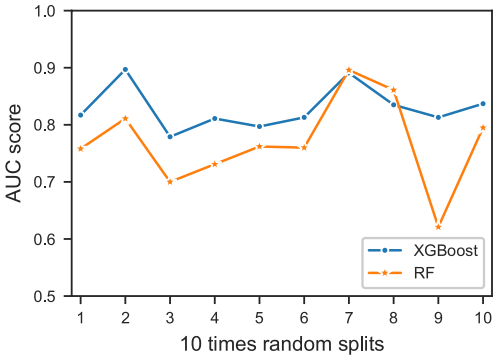
4.4.2 Training and test results for Case No.6

After determining the input parameters of Abutment&Wall dataset, the RF and XGBoost algorithms are applied and the random splits are repeated 10 times.

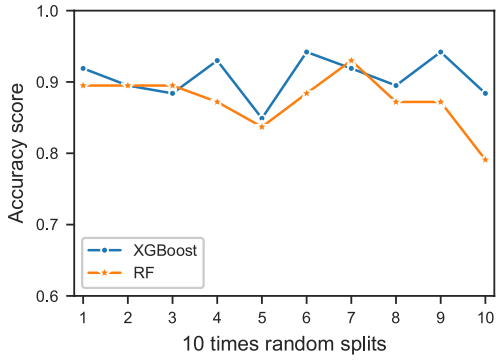
Figure 4.12 presents the training and test scores under 10 times random splits. It is clearly observed that compared with the RF classifier, XGBoost model has higher AUC, accuracy, precision, recall scores and lower FPR. Although random splits are conducted, it's the same data set that is tested by the two algorithms each time. It can be seen that results are more scattered in the RF classifier.



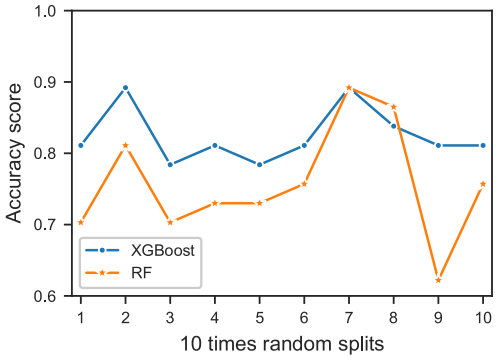
(a) AUC training scores



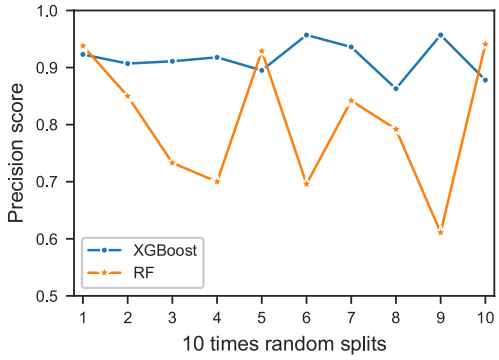
(b) AUC test scores



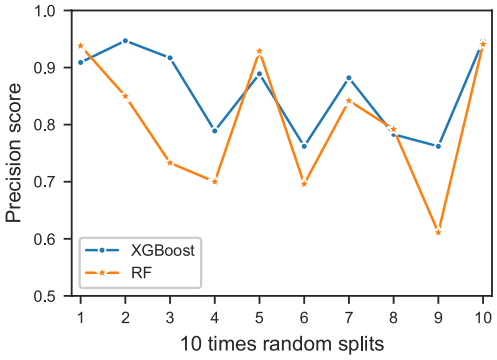
(c) Accuracy training scores



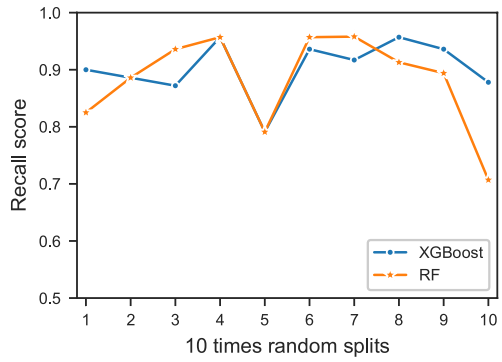
(d) Accuracy test scores



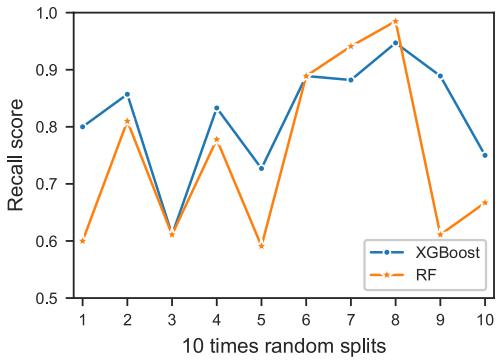
(e) Precision training scores



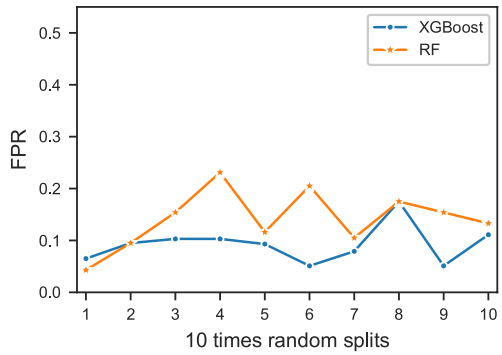
(f) Precision test scores



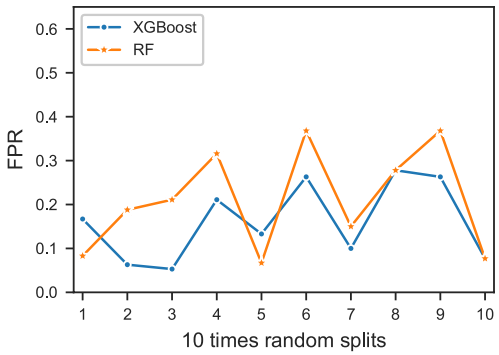
(g) Recall training scores



(h) Recall test scores



(i) FPR training scores



(j) FPR test scores

Figure 4.12: Training and test scores under 10 times random splits (Abutment&Wall dataset)

Later, the average and standard deviation values of model performance measurements for training and test data are summarized in Table 4.8 and Table 4.9 respectively.

Table 4.8: Results on training data over model performance measures for Case No.6 (Abutment&Wall dataset)

	AUC	**Accuracy**	**Precision**	**Recall**	**FPR**
XGBoost	0.905±0.0030	0.906±0.030	0.915±0.031	0.903±0.050	0.092±0.036
RF	0.871±0.038	0.874±0.038	0.877±0.037	0.882±0.084	0.141±0.055

Table 4.9: Results on test data over model performance measures for Case No.6 (Abutment&Wall dataset)

	AUC	**Accuracy**	**Precision**	**Recall**	**FPR**
XGBoost	$\mathbf{0.829 \pm 0.038}$	$\mathbf{0.825 \pm 0.039}$	$\mathbf{0.859 \pm 0.076}$	$\mathbf{0.819 \pm 0.099}$	$\mathbf{0.161 \pm 0.088}$
RF	0.770±0.078	0.757±0.080	0.803±0.116	0.748±0.152	0.211±0.117

Compared with the RF classifier, XGBoost model has higher accuracy, precision, recall, AUC in average and lower FPR. Based on these results, variables in Case No.6 are considered as the most optimal ones for Abutment&Wall dataset and they respectively are:

- **Environmental**: slope of riverbed (CA2), specific flood flow (m^3/s) (CA3), width of valley/width of low flow channel (CA4), flow sinuosity (CA6);
- **Bridge characteristics**: existence of foundation scour countermeasures (BA9), attack angle (BA10);
- **History**: scour history (HA11), flood frequency (HA12);
- **Changing factors**: susceptible of scour (IA13), channel rating (IA14), riverbank rating (IA15), existence of dislocation or deformation around masonry or gabion (IA16), existence of local scour (IA17), rating of other damages of foundation (corrosion, timber piles degradation, cracks, etc.) (IA18).

Similar to bridge pier dataset, it's the XGBoost algorithm that has a better performance for predicting the scour risk of bridge abutment and retaining walls.

However, due to the lack of data, we conduct several tests to determine the optimal variables for Abutment&Wall dataset according to the feature importance rank in Figure 4.10. It can be seen that the the test results in terms of accuracy and other

measurements of XGBoost algorithm are not as good as bride pier dataset, which is possibly caused by the non-sufficient data as well. The process to select input variables and the classifier of Abutment&Wall dataset should be seen as a first try. Once more data is collected, the classifier could probably have a better performance and the excluded variables could be included.

4.5 Conclusions

In this chapter, two machine learning models are constructed for scour risk evaluation: one for bridge piers, and another for bridge abutments and retaining walls. To train the models, several commonly used machine learning algorithms are applied. Machine learning models' performance is evaluated using carefully selected measurements under 10 times random train-test data splits.

Concerning the bridge pier dataset, XGBoost and RF models have the most promising results. The XGBoost classifier achieves high accuracy (0.959/0.938), precision (0.970/0.961), recall (0.974/0.956) and low false positive rate (0.085/0.114) for training and test set respectively. The classifier obtains an AUC score equal to 0.974 for test set (a perfect classifier has an AUC equal to 1). Moreover, XGBoost model provides a feature importance plot: specific flood flow (m^3/s), slope of riverbed (%), existence of dislocation or deformation, existence of local scour were considered as the four most important features. The single feature experimental results imply that it's the interaction among features that generates the scour phenomenon. RF classifier, on the other hand, has a slightly higher recall score, which means it could identify high scour risk piers more accurately than the XGBoost model. But at the same time, the low scour risk pier is more probable to be evaluated as high in RF classifier. Chapter 5 presents a more detailed comparison between these two classifiers.

As for the Abutment&Wall dataset, due to its smaller data size, an overfitting is

observed when training the model at the beginning. After conducting several tests, the XGBoost algorithm with variables in Case No.6 has the best performance. In order to select the optimal features for this dataset, feature importance of the XGBoost classifier is referred to. It should be noted that there are other techniques in the literature to avoid overfitting (introduced in section 2.5.3.2) and they could be tried as well. Despite the various techniques to avoid overfitting, most importantly, it's necessary to collect more data for this dataset. Current size with selected features are difficult to build a machine learning model, let alone to have a robust and reliable one.

The ML models in this chapter suggest that they could be effective and accurate in bridge scour risk evaluation and possibly be employed as an alternative approach in the future. Up to now, step 1-4 shown in Figure 2.20 are realized. The rest of the work will focus on examining the model's reliability, practicability (Chapter 5) and making the model more transparent (Chapter 6).

Chapter 5

Model validation and comparison

5.1 Introduction

The machine learning models' performance is examined by several measurements calculated from a confusion matrix in Chapter 4. Despite that the classifiers we have built achieve encouraging results, they are trained with a limited number of data and it is not quite sure how the model will perform on the unseen data.

In this circumstance, the objective of this chapter is to validate the classifiers built in Chapter 4 by testing them to more cases. A series of analyses are conducted regarding the XGBoost and RF models in which very promising results are obtained for the bridge pier dataset. As a reminder, XGBoost classifier has the best prediction accuracy for both classes and RF model detects better the high scour risk classes because of its high recall score.

To achieve these goals, the two ML models are examined in terms of their robustness and practicality. Furthermore, they are compared with two practical guidelines (scoring table and ARPSA) introduced in Chapter 2. Bridges in Japan and France

are evaluated by ML models and existing approaches. All these efforts allows understanding, and comparing the XGBoost and RF models from different perspectives. In the end, it could help users (inspectors/engineers) choose wisely between them from not only the perspective of prediction accuracy but also robustness, practicability, etc.

This chapter is organized as follows: section 5.2 conducts the parametric study to investigate the robustness of ML models. Section 5.3 compares the prediction results with two engineers having different working experience (entry level and senior level). Comparison between ML models and two existing approaches are shown in section 5.4. In the end, the advantages and disadvantages of XGBoost and RF models are discussed in section 5.5.

5.2 Robustness investigation

Despite machine learning has been applied in numerous studies, only a little number of them have investigated model robustness. In this section, the background and importance of robustness investigation are introduced firstly. Later, the input parameters are reclassified in order to quantify the contribution of parameters. In the end, two types of robustness analyses, namely monotonicity and uncertainty analyses are realized.

5.2.1 Background

Shahin (2013) once pointed out the that besides primary steps to construct a ML model for prediction (e.g., data division, data preparation, model validation), examining the model robustness, which means the predictive ability of ML model to generalize over a range of data similar to that used for model training, is one of the supplementary aspects to enhance model reliability.

Only a limited number of studies in the literature investigated this aspect (Zhang et al., 2020; Zhang et al., 2021a; Wang et al., 2020a; Shahin et al., 2005). Zhang et al. (2021a) did a robustness investigation regarding the clay compressibility. All ML models have a satisfactory performance over the training and test sets. Later in parametric study, by making input parameters except one fixed to mean value and varying the studied parameter from the minimum to maximum values, the soil compression index (C_c) differentiate among algorithms (see Figure 5.1) by changing monotonically the liquid limit water content (w_L). Most importantly, results coming from random forest algorithm fluctuate a lot and are difficult to be justified from a physical understanding. Zhang et al. (2021a) pointed out this fluctuation may be due to the distribution of input parameters and the classification condition at each node in random forest.

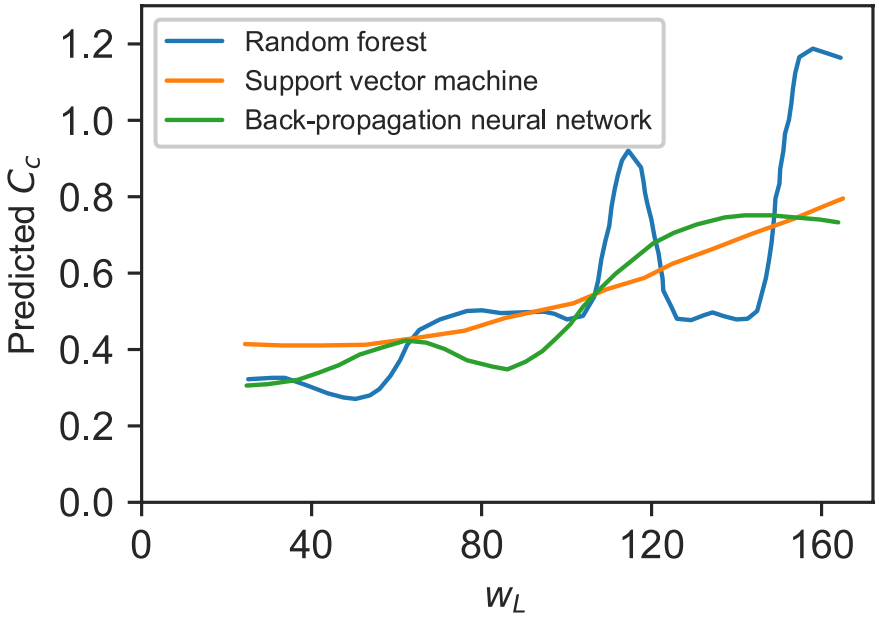


Figure 5.1: Correlation between predicted compression index (C_c) and liquid limit water content (w_L) (Zhang et al., 2021a).

The study of Zhang et al. (2021a) indicates that although the model performs well

against the traditional measurements such as RMSE and R^2, we should also examine how well the predicted output agrees with the known underlying physical process. Once the model is constructed, more comprehensive work needs to be realized to prove the reliability and generality of ML model to make it trustworthy.

5.2.2 Input parameters reclassification

In order to quantify the contribution of model input parameters to prediction results, the original 18 input variables are reclassified according to the theory of Li et al. (2010), in which factors related to natural catastrophic risk can be divided into three aspects: inducing factors (I), environmental factors (E), and vulnerability (V). Below are short descriptions of each aspect:

- **Inducing factors (I)**: they are mainly the direct cause for natural hazard and are linked closely to the occurrence of catastrophic losses. For example, scour induced bridge failures usually directly result from the increase of river flow.

- **Environmental factors (E)**: these factors refer to the environment that breeds the disasters. Such factors play the role to whether mitigate or aggregate the destructive power of natural hazard. For example, a local scour is more likely to be detected in a mountainous area where the slope of riverbed is steep instead of a plain area where the flow is steady.

- **Vulnerability (V)**: the degree to which a system is likely to experience and adapt to harm due to exposure to a natural hazard. Bridge vulnerability regarding scour risk may be its foundation type, construction year, history, and observed damages in the field.

Table 5.1 shows the classification of input variables based on the theory of Li et al. (2010). It should be noted that the inducing factor (I1) and two environmental factors (E3, E4) are quantitative variables, while the rest environmental factors (E2, E5, E6,

E7) and all vulnerability factors (V8, V9, V10, V11, V12, V13, V14, V15, V16, V17, V18) are qualitative ones.

Table 5.1: Input parameters reclassification by adapting the theory of Li et al. (2010)

Category	**Variable(s)**
Inducing factor (I)	Specific flood flow (m^3/s) (I1)
Environmental factors (E)	Flow type (E2), Slope of riverbed (%) (E3), Width of valley/Width of low flow channel (E4), Topography (E5), Flow sinuosity (E6), Riverbed material (E7)
Vulnerability factors (V)	Pier shape (V8), Foundation type (V9), Existence of foundation (V10), Scour history (V11), Flood history (V12), Susceptible of scour (V13), Channel rating (V14), Riverbank rating (V15), Existence of dislocation or deformation around masonry or gabion (V16), Existence of local scour (V17), Rating of other damages of foundation (V18)

5.2.3 Monotonicity and uncertainty analyses

5.2.3.1 Investigation methodology

Monotonicity analysis has been applied in several geotechnical studies to see the robustness and generalization ability of ML model (Zhang et al., 2021a; Shahin et al., 2005; Wang and Yin, 2020). In monotonicity analysis, the investigated parameter varies from the minimum to maximum and other parameters are fixed at mean values.

However, the input variables are all numerical in the aforementioned studies. Regarding the categorical variables in our dataset, it's not easy to determine their mean values. Thus, the fixed values for categorical variables are the subcategory which takes the highest proportion. Instead of varying from the minimum to maximum

value, the investigated categorical variable is varied from each subcategory. For example, the variable flow type (C1)'s fixed value is "*fluvial*", since it's the majority. When it becomes the investigated parameter, the value varies as "*fluvial*", "*other*", and "*torrential*".

The explored range for monotonicity analysis of each variable may refer to Figure 3.6 and Figure 3.7.

Compared with monotonicity analysis which focuses on the performance of ML classifier over the whole dataset, uncertainty analysis tackles this issue in a local scale with concrete examples. It is about the incomplete and uncertain information regarding structure itself or during inspection.

Uncertainty is often divided in two types: epistemic uncertainty, which can be minimized by having more accurate measurements; and aleatory uncertainty, coming from the random nature and cannot be reduced (Kiureghian and Ditlevsen, 2009). For example, ratings of channel or riverbank may be not precise enough due to the incomplete information from the field and this is epistemic uncertainty. With detailed inspection, the value of feature could be more accurate. On the other hand, the flow of watercourse could possibly be changed in the future due to climate change and it is difficult to anticipate the future flood flow due to the aleatory nature of watercourse.

Table 5.2 lists the parameters that may possess epistemic or aleatory uncertainty. In order to observe the performance of the machine learning classifiers, parametric studies are conducted. For quantitative factors, they are considered to be half of or two times larger than the original value while the qualitative factors are varied among all the subcategories.

10 randomly selected bridges located in France are used as case studies. The investigated parameter varies in the given explored range and the rest parameters remain as the original value.

Table 5.2: Selected variables for uncertainty analysis with their explored range

	Variables	**Explored range**
Quantitative factors	Specific flood flow (I1), Slope of riverbed (E3), WV/WC (E4)	0.5 X variable X 2
Qualitative factors	Riverbed material (E7), Susceptible of scour (V13), Channel rating (V14), Riverbank rating (V15), Existence of dislocation or deformation around masonry or gabion (V16), Existence of local scour (V17), Rating of other damages (V18)	All subcategories

In our study, the prediction is in categorical form (high or low scour risk). Supposing the prediction with original feature value belongs to class B, by varying monotonically the investigated variable, three types of possible model performance could happen and it is shown in Figure 5.2.

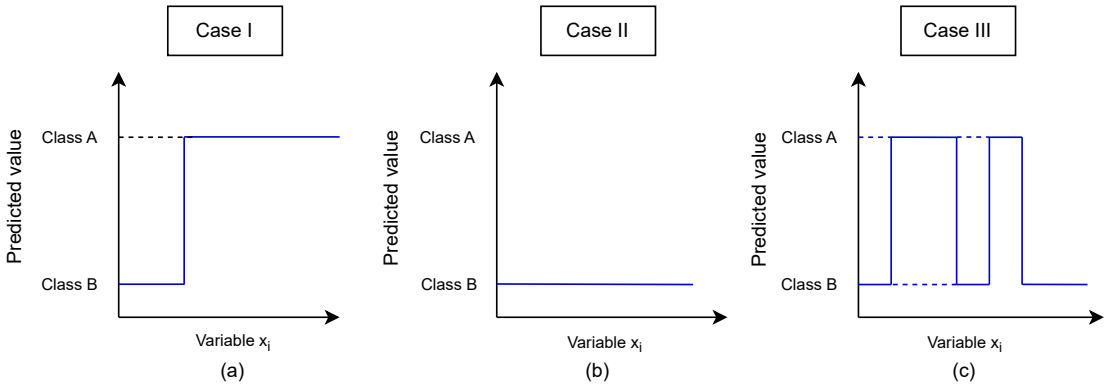


Figure 5.2: Possible model performance after monotonicity or uncertainty analysis.

Case I means that the prediction changes to another class (e.g. from low scour risk to high scour risk). Case II signifies the prediction remains the same. Case III

indicates that the predicted result fluctuates between two classes. Based on physical understanding, if the model performs in the way shown in Case III, the results are surely unreasonable and in conflict with the existing knowledge.

5.2.3.2 Results

- **Monotonicity analysis**

By varying the investigated parameter from the minimum to maximum and other parameters fixed at mean or majority values, Figure 5.3 illustrates two examples of monotonicity analyses results. When changing the variable V11 (scour history) and E4 (width of valley/width of low flow channel), the ML model performs the way as Case I in Figure 5.3: the scour risk is changed low to high (or high to low). Entire results of monotonicity analysis of the two ML models are shown in Table 5.3.

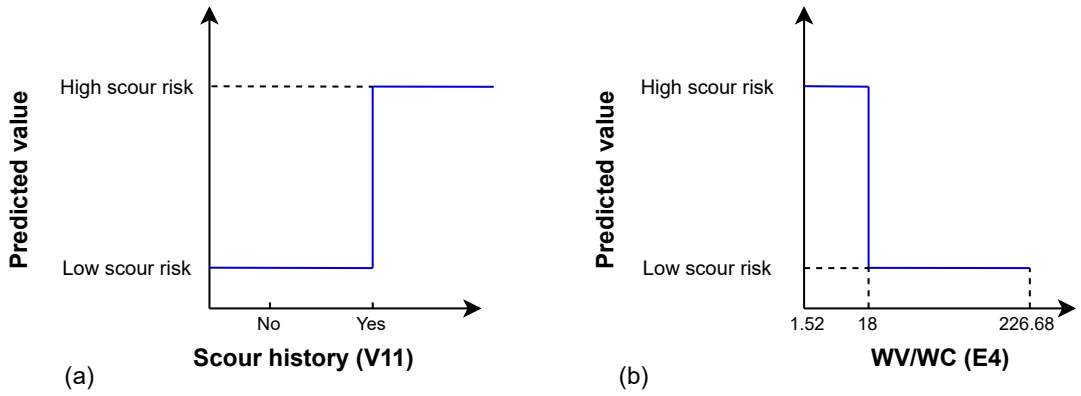


Figure 5.3: Examples of monotonicity analysis for Case I: (a) V11 (Scour history) in RF model; (b) E4 (width of valley/width of low flow channel) in XGBoost model.

Table 5.3: Monotonicity analysis results

	RF	**XGBoost**
Case I	V11, V7	E4, V16, V17
Case II	I1, E2, E3, E4, E5, E6, E7, V8, V9, V10, V12, V13, V14, V15, V16, V18	I1, E2, E3, E5, E6, E7, V8, V9, V10, V11, V12, V13, V14, V15, V18
Case III	-	-

It can be seen that neither of the two models performs in the way shown in case III. In most cases, the predictive class remains the same by changing the studied parameter. However, it's noteworthy that variables causing Case I are not exactly the same between the two algorithms (V11 and V7 for RF model but E4, V16 and V17 for XGBoost model).

- **Uncertainty analysis**

Monotonicity analysis only permits seeing the model performance in a global scale over the whole range of data. The uncertainty analysis, on the other hand, makes it possible to see how the ML model performs by considering data uncertainty. Only variables whose information could be uncertain is investigated. Table 5.4 shows the uncertainty analysis results by using the 10 selected examples.

Table 5.4: Uncertainty analysis results

Case No.	Region	Model	Case I	Case II	Case III
1	Bordeaux	RF	V16	I1,E3,E4,E7,V13,V14,V15, V17,V18	-
		XGBoost	V16	I1,E3,E4,E7,V13,V14,V15,V17,V18	-
2	Lyon	RF	-	I1,E3,E4,E7,V13,V14,V15,V16, V17,V18	-
		XGBoost	V13, V17	I1,E3,E4,E7,V14,V15,V16,V18	-
3	Tours	RF	-	I1,E3,E4,E7,V13,V14,V15,V16, V17,V18	-
		XGBoost	E3	E4,E7,V13,V14,V15,V16, V17,V18	I1
4	Toulouse	RF	-	I1,E3,E4,E7,V13,V14,V15,V16,V17,V18	-
		XGBoost	V16	I1,E3,E4,E7,V13,V14,V15,V17,V18	-
5	Clermont Ferrand	RF	-	I1,E3,E4,E7,V13,V14,V15,V16, V17,V18	-
		XGBoost	-	I1,E3,E4,E7,V13,V14,V15,V16,V17,V18	-
6	Nante - Rennes	RF	V17	I1,E3,E4,E7,V13,V14,V15,V16,V18	-
		XGBoost	V17	I1,E3,E4,E7,V13,V14,V15,V16,V17,V18	-
7	Metz-Nancy	RF	E3, V18	I1,E4,E7,V13,V14,V15,V16,V17	-
		XGBoost	I1,E3	E4,E7,V13,V14,V15,V16,V17	-
8	Strasbourg	RF	V17	I1,E3,E4,E7,V13,V14,V15,V16,V18	-
		XGBoost	V17	I1,E3,E4,E7,V13,V14,V15,V16,V18	-
9	Montpellier	RF	V17	I1,E3,E4,E7,V13,V14,V15,V16,V18	-
		XGBoost	V17	I1,E3,E4,E7,V13,V14,V15,V16,V18	-
10	Montpellier	RF	V16,V17	I1,E3,E4,V13,V14,V15,V18	E7
		XGBoost	I1,V16,V17	E3,E4,E7,V13,V14,V15,V18	-

Table 5.4 indicates that similarly to monotonicity analysis, in most cases the predictive class remains as the original value (Case II) by modifying the investigated parameter each time. However, both algorithms perform in an unrobust fashion (Case III) in the given examples (Case No.3 for XGBoost and Case No.10 for RF), while Case III is not observed in monotonicity analysis.

The number of variables in Table 5.4 in each case is counted by classifying as inducing (I), environmental (E) and vulnerability (V) factors. Results are shown in

Figure 5.4. Figure 5.4 (a) indicates that compared with the RF model, it is easier for XGBoost classifier to change the prediction results. For example, the change of predicted class (Case I) is caused by inducing factor (I) two times in XGBoost classifier while none for RF model. The modification of environmental (E) and vulnerability (V) factors have the same effect: the predictions are more probable to change in XGBoost model than in RF. Seeing from the randomly selected 10 case studies, the prediction results of RF are relatively more stable (remain as Case II) even changing the input parameters, as shown in Figure 5.4 (b).

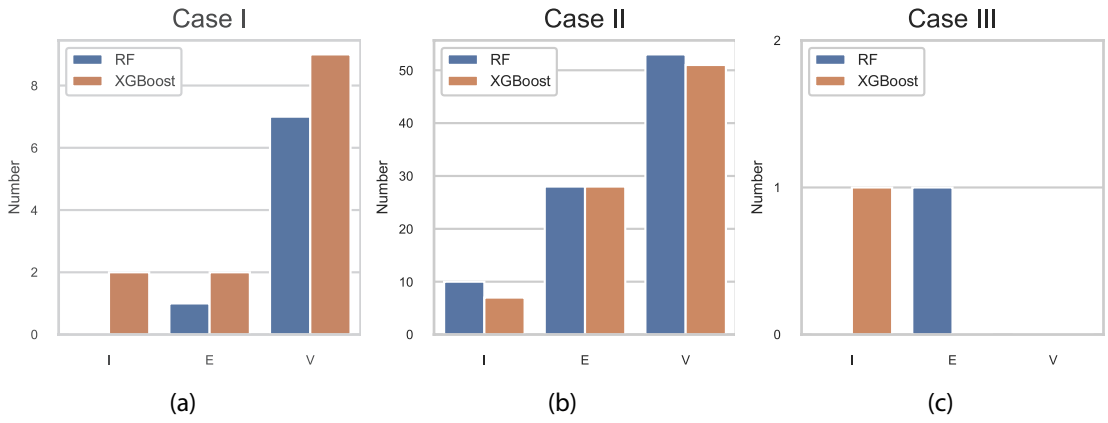


Figure 5.4: Uncertainty analysis results by classifying variables as inducing (I), environmental (E) and vulnerability (V) factors for Case I (a), Case II (b), and Case III (c).

The above analyses compare the RF and XGBoost models by conducting two robustness analyses, with the aim of investigating whether they perform in a reliable and reasonable fashion. While monotonicity analysis focuses on a global scale, the uncertainty analysis is conducted in a local scale with concrete examples. Three types of possible model performance is distinguished. Generally speaking, in most cases the predictive class remains the same by exclusively changing one parameter each time, which seems reasonable since a high scour risk scenario is often triggered by multiple aspects. For example, an increasing flow easily comes with the transportation of de-

bris and the riverbank erosion. In uncertainty analysis, both algorithms perform in a way that is contradictory to the existing knowledge. Such phenomenon has also been observed in other studies (Zhang et al., 2021a; Wang et al., 2020a): the prediction results don't form a smooth curve when using the tree-based algorithm and fluctuate a lot. This phenomenon may result from the discrete distribution of input variables, which ultimately influences the classification conditions at each node.

To conclude, after robustness analysis, we observed that both classifiers can perform in a robust fashion in most cases. To understand why the variables resulting in Case I, Case II and Case III are different in two models relates to the algorithm itself and how the model is built, which is considered beyond the scope of this study. Detailed comparison of XGBoost and RF classifiers by considering the results of robustness analysis is shown in section 5.5.

5.3 Practicality investigation

As presented in Figure 2.1, the machine learning model serves as a complement to the engineer's judgement with the purpose of optimizing the decision-making process. In this section, the RF and XGBoost models are tested by bridges in Occitanie region in France. The bridges are then evaluated by two SNCF engineers' with different working experience. The evaluation results are then compared among different approaches for examining whether the ML model can help the decision-making in practice or not.

5.3.1 Bridges in Occitanie region for testing

In order to examine the practicability of two ML models, they were applied to 40 bridge piers located in the Occitanie region in France. Figure 5.5 shows the pictures of some tested bridges. Bridges tested here have the same construction techniques as the ones in the training and test data sets. They are also constructed almost 100

years ago.

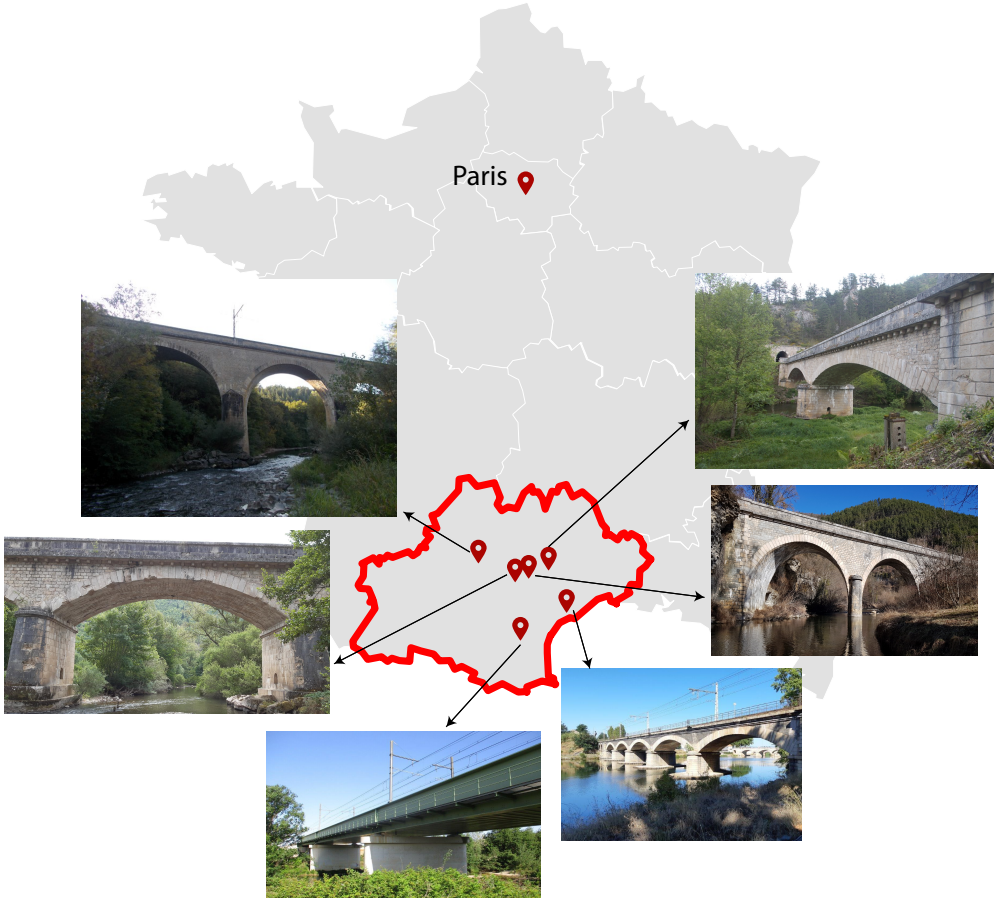


Figure 5.5: Examples of bridges tested in Occitanie Region in France.

Besides evaluating the bridges by applying the ML models, a junior engineer, whose working experience is less than two years, and a senior engineer who works more than five years in the related domain were invited to assess the same elements as well. In the end, the assessment results are compared to see the coherence between ML models and engineers with different working experiences.

5.3.2 Practicality investigation results

The number of the same assessment results among RF, XGBoost classifiers and two engineers is shown in Figure 5.6. It can be seen that they are generally in agreement. The least match occurs between the RF model and the junior engineer but there are still 32 cases (80%) in common. Supposing that the results given by the senior engineer are all "correct", XGBoost model is the one that most close to the senior engineer (38 cases or 95% in common), followed by the junior engineer (37 cases or 92.5% in common) and then RF model (34 cases or 85% in common).

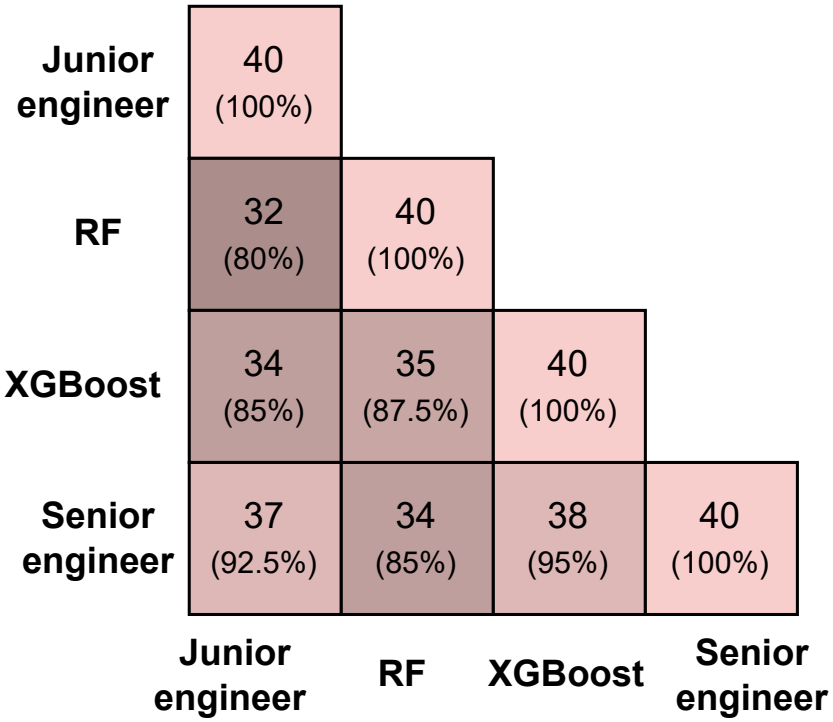


Figure 5.6: Number of the same assessment results among the junior engineer, RF classifier, XGBoost classifier and the senior engineer.

The high coherence between the engineers and the ML models proves the practicability of the proposed ML models. Most importantly, XGBoost classifier is the

one that is the most close to the senior engineer's evaluation. In other words, this test proves the possibility to use the proposed ML based solution in practice as an alternative approach, especially for helping engineers at entry level.

5.4 Comparison with existing approaches

In order to validate the constructed machine learning models, it's important to compare them with existing approaches. In this section, the RF and XGBoost models are compared with the Japanese scoring table and the French ARPSA methods. It is well acknowledged that both scoring table and ARPSA are not built for the French railway bridges. The reason for doing this model comparison is to understand the differences among each approach, which could ultimately help for model improvement. EX2502 is not used for comparison because an overestimation of general scour is observed. Scoring table and ARPSA are introduced in section 2.4.

5.4.1 Scoring table (Japan)

Scoring table is the practical guideline for the Japanese railway bridges' scour risk evaluation. In order to compare the ML models with the scoring table, a methodological comparison is conducted firstly. Later, the XGBoost and RF models are applied to 20 Japanese cases, which are also evaluated by the scoring table.

5.4.1.1 Methodological comparison

Before applying the two procedures, a methodological comparison is conducted at first to see their similarities and differences between scoring table and the ML model. They are concluded and shown in Table 5.5.

Table 5.5: Summary of methodological similarities and differences between the scoring table and the French machine learning model

Similarities
1. Same objective: assess scour risk of bridge element in a practical way.
2. Risk classes as an outcome, and in a binary form.
3. Don't calculate local scour depth, general scour depth with design flood level to conclude risk level.
Differences
1. Foundation depth plays an important role in Japanese approach, but it is not directly included in the ML method due to the difficulty for accessing data. However, variables in ML model such as the bathymetry evolution since the year of construction, and types of damages, in some ways reflects this information.
2. Parameters for describing hydrology and hydromorphology are not included in Japanese method, because information is considered already covered in topographical landform, riverbed materials, and hydraulic conditions.
3. The influence of riverbed particle to scour is different. Japanese model doesn't include cohesive soil (e.g., silt, clay), which is commonly seen materials in the French river. Cohesionless soil (e.g., sand) is considered as the material most likely to increase scour risk in French method.
4. The French ML model comprises the history of the structure (e.g., scour history, flood history) while the Japanese procedure doesn't.
5. In Japanese procedure, the hydraulic structure in the vicinity is an important factor and the bridge can directly be considered at high risk while the ML model doesn't take into account of this feature.
6. The protection condition is scored differently in Japanese guideline based on the scour countermeasure types. The French ML model considers the different scour countermeasures having the same influence.

The nature of the ML model and scoring table is not the same: one is a data-driven approach, and another is calibrated by engineers' experience. Nevertheless, the common objective of the two guidelines is to screen high scour risk in an effective way. Bridges evaluated at high risk will need a detailed inspection, reinforced surveillance, or completed geotechnical and hydrological studies.

However, it can be observed from Table 5.5 that the parameters required in each guideline are quite different, which is primarily caused by the ways of managing the railway infrastructure. In France, when the water level is high, an underwater foundation inspection through diving is conducted. However, the impact vibration test is often adopted in Japan to know the status of bridge piers through comparing the natural frequency. The second reason is the different hydrological conditions. From a general view, compared with Japan, watercourse in France is more stable in terms of velocity and river diversion. Therefore, it causes the differences for riverbed material, the impact of riverbed material to scour and the important role of hydraulic structures (e.g., weir, dam) near bridges.

After the methodological comparison, the two ML models and the scoring table are applied to 20 cases for testing.

5.4.1.2 20 Japanese cases for testing

Similarly to France, Japan also possesses a large number of historical railway bridges. The first Rail Construction Act started in 1892 in Japan for conventional lines (Shibayama, 2017). The bridge construction techniques in conventional rail lines (equivalent to TER in France) are also very similar to France like timber pile foundation, caisson. Recently, scour has been reported as a serious issue in both two countries due to the increased frequency and intensity of extreme weather events (Takayanagi et al., 2019).

The two RF and XGBoost models are trained and tested by the French data. In

order to examine the two models' generality, which means their predictive capacities over bridges outside France, the two classifiers are tested by 20 railway bridges located in Japan. The assessment results are then compared with the scoring table (ST). Testing the French ML models using Japanese data and then comparing with the Japanese guideline could, in a way, validate the two ML models as well.

Data from the 20 Japanese cases were provided by the Railway Technical Research Institute (RTRI). Their locations and photos of some bridges are shown in Figure 5.7. Among these cases, JP1-JP10 are bridges that once experienced severe flood events (e.g., after typhoon or heavy rain) while JP11 – JP20 are the ones without obvious damages in history. The 10 damaged cases (JP1-JP10) comprise information before and after flood events, so they are assessed at these two moments as shown in Figure 5.8. Figure 5.7(c) shows the photo of JP5 after flood event (AF).

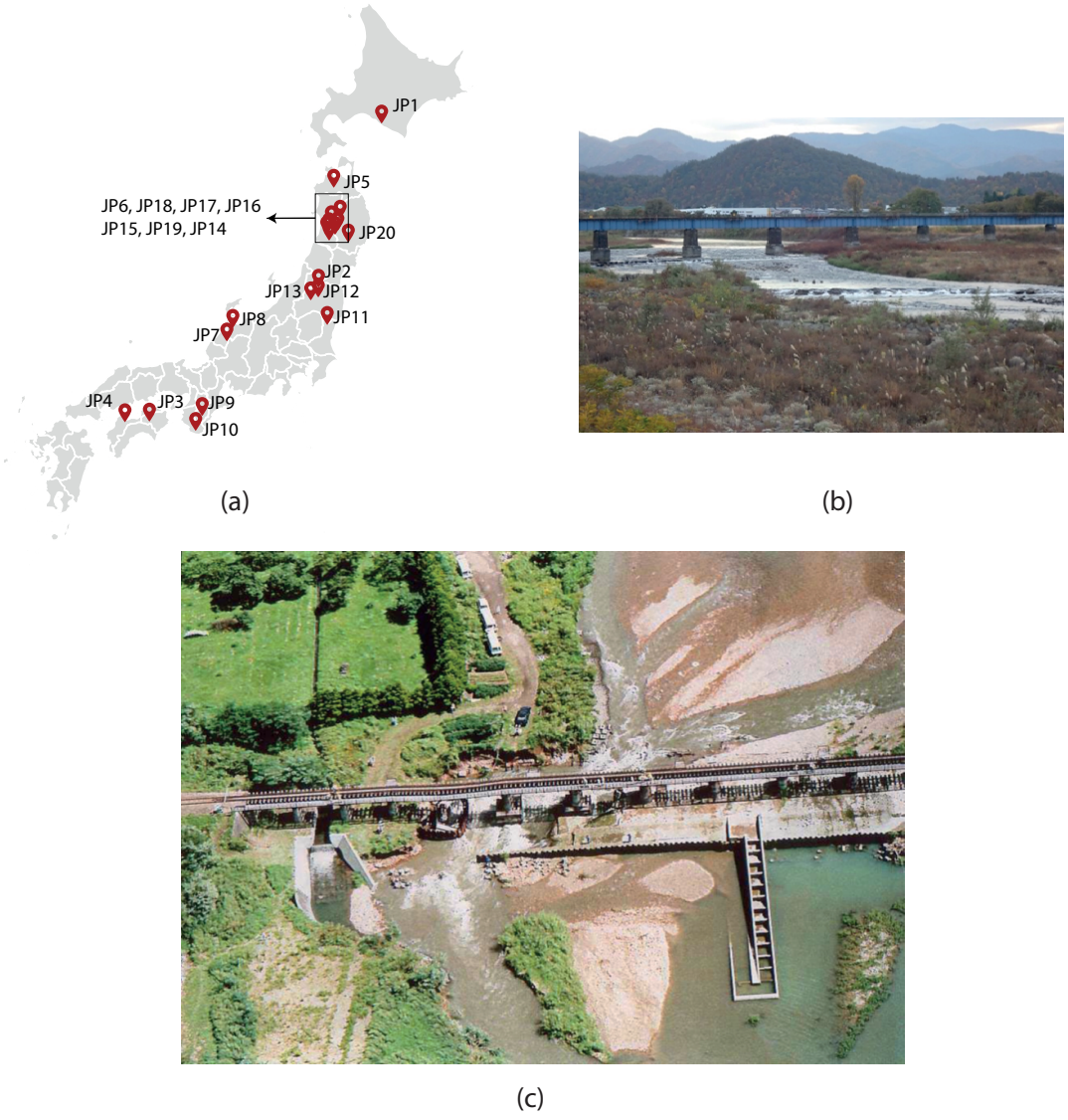


Figure 5.7: Case studies of bridges in Japan: (a) locations of 20 bridges; (b) photo of JP2 after foundation regeneration; (c) photo of JP5 after the flood event (Samizo, 2014).

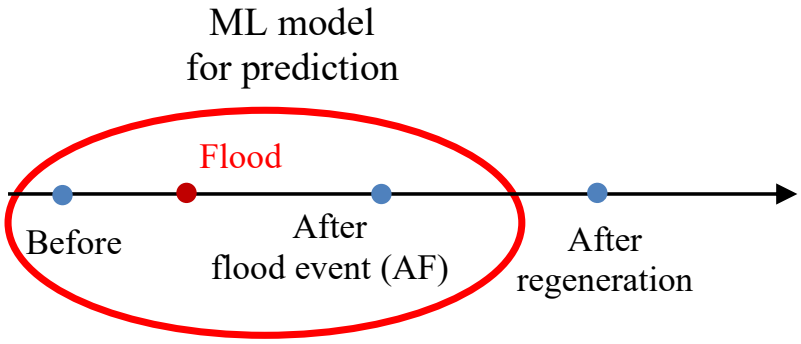


Figure 5.8: Timeline of the ten damaged bridge cases.

5.4.1.3 Japanese cases test results

20 Japanese bridges are tested by the trained RF, XGBoost classifiers and the Japanese guideline scoring table (ST). Since JP1 – JP10 are tested before and after the flood events, there are 30 cases in total for testing. Very similar to France developed ML classifiers, the output of ST is also in a binary form, namely low scour risk and high scour risk, which allows comparing directly the results obtained from ML classifiers without model calibration.

Table 5.6 shows the assessment results of Japanese cases via different approaches. It can be observed that the same case has quite a different predicted class as output among all approaches. Most cases are assessed at high risk by using RF classifier and ST (20 and 24 respectively). On the contrary, 19 cases are evaluated at low scour risk when applying XGBoost model, compared with only 6 and 10 from ST and RF classifier respectively.

Table 5.6: Assessment results of 20 JP cases tested in XGBoost, RF and ST

Class	**Approach**	**Case No.**	**Total count**
Low scour risk	RF	JP6, JP9, JP12, JP14, JP15, JP16, JP17, JP18, JP19, JP20	10
	XGBoost	JP1, JP2, JP3, JP4, JP4AF JP5, JP7, JP7AF, JP8, JP9, JP11, JP13, JP14, JP15, JP16, JP17, JP18, JP19, JP20	19
	ST	JP1, JP11, JP12, JP15, JP16, JP17	6
High scour risk	RF	JP1, JP1AF, JP2, JP2AF, JP3, JP3AF, JP4, JP4AF, JP5, JP5AF, JP6AF, JP7, JP7AF, JP8, JP8AF, JP9AF, JP10, JP10AF, JP11, JP13	20
	XGBoost	JP1AF, JP2AF, JP3AF, JP5AF, JP6, JP6AF, JP8AF, JP9AF, JP10, JP10AF, JP12	11
	ST	JP1AF, JP2, JP2AF, JP3, JP3AF, JP4, JP4AF, JP5, JP5AF, JP6, JP6AF, JP7, JP7AF, JP8, JP8AF, JP9, JP9AF, JP10, JP10AF, JP13, JP14, JP18, JP19, JP20	24

To evaluate the model similarity, Figure 5.9 shows the number of the same assessment results among the three evaluation methods. RF model is the one most close to ST with 22 out of 30 (73%) cases in common while for XGBoost classifier there are 15 cases (50%).

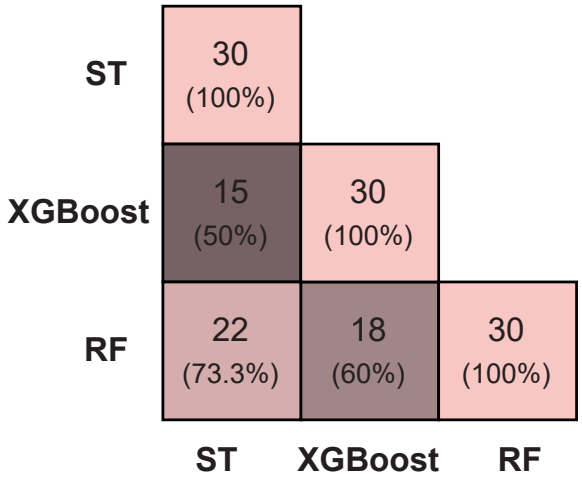


Figure 5.9: Number of the same assessment results among ST, XGBoost and RF.

Concerning the Japanese cases, instead of the XGBoost classifier, it's the RF model which is closer to the Japanese guideline. The frequency and intensity of natural hazards (e.g., storm, flooding, heavy rain) in Japan are much higher and more severe than in France. Consequently, a bridge evaluated at a low scour risk level by French standards could be in a more vulnerable state by following the Japanese guideline, which is built under the environmental conditions in Japan. RF model, on the other hand, has higher recall and false positive rate (see Table 4.2), which means low scour risk cases are more probable to be considered as high. That's the reason why RF model is closer to the results given by ST.

It should be noted that ST is an empirical approach established on case studies. Extra efforts such as confirmation from Japanese expert(s), tested by other approaches are needed to validate the results from ST, which is considered beyond the scope of this work.

Another noteworthy point is that both XGBoost and RF models are trained by using a French dataset. 20 bridges as case studies are not enough to prove the proposed

ML model is ready to be applied in Japan despite the encouraging results. The objective of this test is to demonstrate the possibility of proposed ML model being used in a broader geographical background. To adopt it in practice, the classifier must be trained by using data from the country.

5.4.2 ARPSA (France)

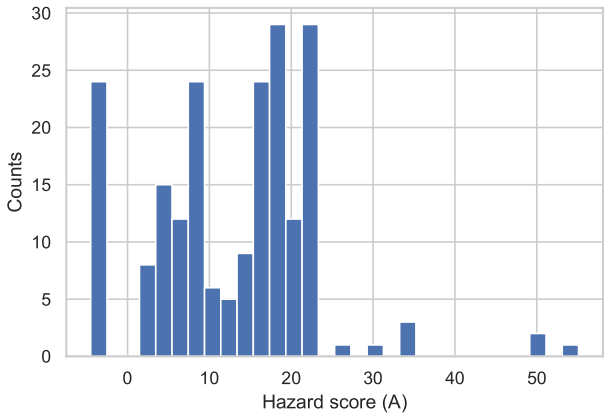
ARPSA is the French scour risk evaluation procedure for highway bridges. There is a qualitative analysis at first in ARPSA. Bridges assessed at medium or high risk are required to have a quantitative analysis then.

Compared with railway bridges, highway bridges are constructed much later with advanced techniques. However, in several cases, highway and railway bridges are often near to each other. That is to say, they share very similar geographical and hydrological environment. Therefore, by applying ARPSA to railway bridges, it allows seeing how the railway bridges are judged by the highway standard.

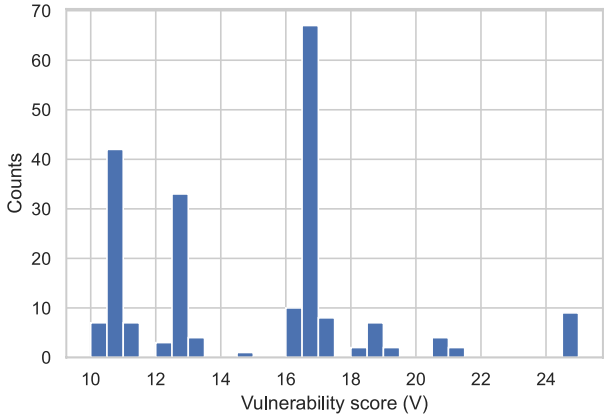
ARPSA consists of two steps' analyses: qualitative (step 1) and semi-quantitative (step 2) analyses. We use at first the whole bridge pier dataset for qualitative analysis. Later, two bridges are selected to conduct the quantitative analyses. The selected two bridges are then compared with the RF and XGBoost models.

5.4.2.1 Step 1: qualitative analysis

The whole bridge pier dataset is used for conducting the qualitative analysis, in order to see the percentage of bridges requiring the semi-quantitative analysis (step 2). The hazard and vulnerability scores are calculated according to Figure 2.12 and Figure 2.13. The score histograms are shown in Figure 5.10.



(a)



(b)

Figure 5.10: Histograms of hazard (a) and vulnerability (b) scores.

Since ARPSA is designed for highway bridges, to validate whether we have adopted it correctly on railway bridges, the minimum (Min.) and maximum (Max.) hazard and vulnerability scores obtained in our dataset are compared with the work of Younsi (2019). Younsi (2019) once used ARPSA to evaluate 12 railway bridges, which are only located in the region of Provence-Alpes-Côte d'Azur in south of France. The comparison results are shown in Table 5.7.

Table 5.7: Hazard and vulnerability scores compared with the study of Younsi (2019)

	Hazard score (A)		**Vulnerability (V)**	
	Min.	**Max.**	**Min.**	**Max.**
Study of Younsi (2019)	13.2	25.3	14.5	24.5
Present study	-4.5	41.5	10	25

From Table 5.7, it is observed that the vulnerability scores we have calculated are generally in the same range as the study of Younsi (2019). However, the range for hazard score is much larger in our study. This is because vulnerability is the weakness of the structure when facing the natural events. Since in both studies, the objects are railway bridges, who share the same construction techniques and structural types, the range for vulnerability score could be almost the same. Hazard, on the other hand, is defined as the condition (flood event) which could possibly cause the undesirable events. The study of Younsi (2019) only focused on one region in France and that is to say, the geographical background is not very diversified. In our dataset, bridges are in almost every region in France (see Figure 3.3). It's pretty natural that the hazard scores are in a wider range.

The negative hazard scores in Figure 5.10a result from the riverbed material. When the riverbed material is rock, the hazard value could easily be negative. Hazard score larger than 25.3 is because several bridges crossing a torrential rivers and the riverbed material is cohesive soils (silts, clay) or sand. In this case, the general scour (A1) could be 23 or 30 (see Figure 2.12). However, when the bridges in the study of Younsi (2019) cross the torrential river, the riverbed material is always gravel (which means the general scour A1 is equal to 11).

Hazard and vulnerability class can be obtained by following the rules shown in Table 2.5 and Table 2.6. Then, the criticality level, which is the combination of

hazard and vulnerability levels as shown in Table 2.8, is determined.

Table 5.8 shows the number of vulnerability, hazard, and criticality levels after qualitative analysis. Originally, to conclude the risk level in ARPSA, the consequence level is required (see Table 2.9). However, the consequence level classification rule is one the basis of highway standard, and they are not adaptable for railway bridges. Thus, the criticality level is used to decide whether a semi-quantitative analysis (step 2) is needed. Here, we follow the strategy adopted in the study of Younsi (2019)): bridge at medium or high criticality level should have a quantitative analysis.

Table 5.8: Application of ARPSA for qualitative analysis (step 1)

	Vulnerability	**Hazard**	**Criticality**
Low	0	24	5
Medium	59	35	25
High	149	149	178
Sum	208	208	208

From Table 5.8, it is found that most bridges in our dataset require a quantitative analysis: 25 at medium criticality level and 178 at high. Thus, in the next subsection, two bridges are selected as case studies for quantitative analysis and the results are compared with the machine learning model.

5.4.2.2 Step 2: semi-quantitative analysis

Figure 5.11 shows the photos of two bridges that we have selected as case studies.

(a)

(b)

Figure 5.11: Photos of Richebout bridge (a) and Viaduc sur l'Adour (b)

Richebout bridge (Figure 5.11a) is located in Butry-sur-Oise, a city in the north of Paris in France. It crosses the Oise River which flows into the Seine River. Richebout bridge is used for rail and road circulations. The bridge was constructed in 1915 and then reconstructed after World War II. It consists of three spans with a steel deck. The two masonry bridge piers are in river channel. According to the archives, the bridge pier foundations were reinforced in 1981. The velocity of river is measured

between 0.3 to 0.5 m/s during inspection, and the slope of riverbed is around 0.01%. According to recent inspections in 2016 and 2020, no obvious damage is observed in Richebout bridge.

Viaduc sur l'Adour (Figure 5.11b) was constructed in 1881. It's in Saint-Sever commune, Nouvelle-Aquitaine administrative region in southwestern of France. The viaduct crosses the Adour river, which flows into the Atlantic Ocean. The bridge pier (P1), which is in the low flow river channel, is built on caisson foundation. By comparing with photos taken in 1950, it was observed that the width of low flow channel decreased a lot due to the agricultural reason and a quarry in the upstream. The river bed has moved towards pier P1. Inspection in 2022 reported an obvious lowering of riverbed. Moreover, the foundation embedment depth has decreased 4.6 m since its construction (embedment depth were 5.22 m in 1881 and 0.62 m in 2021). Given to these observations, more frequent surveillance regarding this Viaduct and the reinforcement work for pier P1 are required.

The two bridges introduced above are evaluated by ARPSA, XGBoost and RF models. Results of qualitative (step 1 in ARPSA) are firstly shown in Table 5.9.

Table 5.9: Qualitative analyses results (step 1)

	Hazard level (score)	**Vulnerability level (score)**	**Criticality**
Richebout bridge	High (7.1)	High (20.5)	High
Viaduc sur l'Adour	Medium (5.6)	High (12.5)	High

Both bridges have the high criticality level in step 1, which means the semi-quantitative (step2) analysis is required and the results are shown in Table 5.10.

Table 5.10: Semi-quantitative analyses results (step 2)

	Hazard level (score)	**Vulnerability level (score)**	**Criticality**
Richebout bridge	High (6.3)	High (13)	High
Viaduc sur l'Adour	High (7.3)	Medium (10)	High

In semi-quantitative analysis, Richebout bridge is at high hazard and vulnerability level. Consequently, the criticality level for Richebout bridge is high in the end. Viaduc sur l'Adour also has the high criticality level, despite the medium level for vulnerability.

The two bridges are then tested by the ML models, and the predicted scour risk levels are shown in Table 5.11.

Table 5.11: ML model results of Richebout bridge and Viaduc sur l'Adour

	Scour risk	
	XGBoost	RF
Richebout bridge	Low	Low
Viaduc sur l'Adour	High	High

Viaduc sur l'Adour is evaluated at high criticality level in ARPSA, and high scour risk level in two ML models. However, for Richebout bridge, it is evaluated at high criticality by ARPSA while the predictions from XGBoost and RF models are both low scour risk levels.

The hazard and vulnerability levels in ARPSA are determined by the corresponding scores and predefined threshold (examples in Table 2.5 and Table 2.6). When applying ARPSA to railway bridges, we observed that the vulnerability level in both step 1 and step 2 analyses require the construction periods (see Figure 2.13) and the construction

period defined in ARPSA are not completely adaptable for railway bridges: bridges constructed before 1950 have a score equal to 5. Most railway bridges are constructed before this date. In the end, by following the vulnerability classification level, it is pretty natural that the tested railway bridge is at high vulnerability level. In other words, when judging by the road standard, railway bridges are highly possible to have a high vulnerability level and consequently, a high criticality (or risk) level. That's the reason why, despite the fact that no obvious damage is observed in Richebout bridge, ARPSA evaluates Richebout bridge at a high criticality level while both ML models predicts at low.

To conclude, after applying ARPSA to the bridge pier dataset, it is found that most bridges require a semi-qualitative (step 2) analysis. ARPSA could overestimate the scour risk of the railway bridge because for the road standard, most railway bridges are at a vulnerable state due to its construction periods. In order to make ARPSA adaptable for railway bridges, the construction periods and the corresponding classification rules should be modified.

5.5 Discussion

After completing a series of investigations and model comparisons, the advantages and disadvantages of RF and XGBoost models are discussed in this section.

It is observed in Chapter 4 that RF classifier is more sensible to high scour risk cases because of its high recall score. However, the price is that when adopting RF classifier, it could generate high maintenance costs, since low scour risk cases care more probable to be evaluated as high. Thus, we think RF model is more adaptable for regions where the frequency and intensity of natural hazards are high. It could also be applied to regions where rail line serves as a major transportation because in these places, bridge failures could cause severe social disruptions. We must avoid

the traffic disruption and the high recall score of RF classifier could help achieve this objective by detecting high scour risk cases more accurately.

XGBoost classifier has a good predictive capacity for both classes because it has the highest accuracy score among the four ML models. It's also the model that is most closely to the senior engineer's evaluation results. Nevertheless, it's capacity to detect high scour risk classes is not as good as RF. Therefore, XGBoost classifier is suggested to be applied to regions where the inundation risk is low, because the probability to have scour induced bridge failure in these places could be low. It can also be applied to regions who have a constrained budget for maintenance or there are alternatives for transportation besides rail. Moreover, the uncertainty analysis results reveal that the XGBoost classifier requires the data with less uncertainty. If not, the predictive class is more easily to be changed.

5.6 Conclusions

Different types of analyses are conducted in this chapter with the aim of validating and comparing the XGBoost and RF classifiers, which are built in Chapter 4 for the bridge pier dataset.

The robustness investigation allows examining the model performance over unseen data. It is observed that in most cases, the two models can perform in a reliable fashion. The XGBoost model is more sensible to the uncertainty of data because the predicted class is more easily to be changed from one to another.

Later, the two ML models are compared with two engineers having different working experience for practicality investigation. The high coherence between the engineers and the ML models is observed. Most importantly, XGBoost model has 95% in common with the predictions from the senior engineer.

In the end, the ML models are compared with two existing practical guidelines (scoring table and ARPSA) by applying them to Japanese and French bridges. It is found that RF classifier is closer to the scoring table due to its high recall score.

When applying ARPSA to the railway bridges, it is found that although railway and highway bridges could often share similar geographical and hydrological background, ARPSA is not completely adaptable for railway brides. It could overestimate the scour risk because it considers most railway bridges are already in a very high vulnerability level.

After a series of analyses, we think both RF and XGBoost classifiers are capable of being applied in practice. However, RF model could be preferred by region where the frequency and intensity of natural hazards are high or rail line serves as a major transportation. XGBoost model, on the other hand, could be applied to regions where the inundation risk is low or there exists alternatives for transportation besides rail.

The robustness, and practicality investigations in this chapter permit understanding well the two ML models besides the model performance measurements we chose in Chapter 4. Moreover, comparing ML models with existing approaches provides insights for other countries' transport agencies who want to develop their ML based maintenance policy for managing the infrastructure affected by floods.

Given its advantages and disadvantages, XGBoost classifier is chosen as an example for model interpretation and implementation in practice, which are going to be presented in Chapter 6.

Chapter 6

Model interpretation and implementation

6.1 Introduction

In previous chapters, we have built the ML classifiers for prediction (Chapter 4). The practicality, robustness of the ML models have been examined (Chapter 5). However, when looking at Figure 2.20, it highlights the importance to have a transparent model. Therefore, the objectives of the last chapter in this thesis are to interpret the black box ML model, make it transparent by using the explainable artificial intelligence (XAI) techniques. Furthermore, the engineering application is presented at the end of this chapter. A web site is constructed by using the research outcome of this study, with the aim of facilitating the use of ML model in practice.

6.2 ML model interpretation using XAI and engineers' expertise

In this section, XAI approaches and engineers' expertise are employed for interpreting the XGBoost classifier built for the bridge pier dataset.

6.2.1 Importance of building an explainable ML model

When looking at the literature, although researchers keep widening the boundaries of what can be learnt by machine learning algorithms, little has been implemented (Naumets and Lu, 2021), especially in the field of civil engineering.

Machine learning model is also known as black-box model. From this name, it can be understood that the opaque nature of learning algorithms prohibits having a straightforward explanation of the prediction results (Emmert-Streib et al., 2020). For example, compared with a linear regression model, a deep learning neural network may achieve higher accuracy but is difficult to be interpretable. This non-transparency nature makes policymakers reluctant to embrace AI/ML techniques in practice considering the social and legal liabilities taken by civil engineers. Questions may be asked as follows: (1) How has been each prediction made ? (2) To which degree to trust a model's prediction ? (3) Can we challenge the prediction results ?

Scour risk is evaluated after engineer's field inspection. A ML-based model without explanation may not be easily accepted by field engineers who are considered as end-users. Firstly, scour is related directly to bridge failure. In such high risk project, a misclassification may possibly threat human's life and the surrounding environment. Moreover, civil engineers are trained to follow norms and standards. They get used to procedures with explicit explanations. The black-box ML model is nonetheless not capable of telling how each prediction is made. Engineers are consequently reluctant to trust the prediction even though satisfactory test results can be achieved some-

times. Lastly, even though there exits systematically a series of metrics to evaluate the performance of ML model, these measurements are seemingly only the interest of research community (Barredo Arrieta et al., 2020). Although this may be fair for some disciplines, society and science care way more than just performance. Obtaining good results with selected measurements does not effectively indicate that the mechanism behind the problem is revealed but rather infers the specific algorithm could predict well in the established dataset, data from which is presumably considered being representative enough to the phenomenon.

In recognition of the obstacles to promoting AI/ML techniques from academia to industry, the concept of explaining the black-box model, namely explainable AI (XAI) has emerged recently. Although talking about what is explanation enters the realm of philosophy (Díez et al., 2013) and is considered beyond the scope of this paper. To shed some light, Gunning (2017) defined XAI as:
"a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."
In other words, XAI tries to help human users understand the prediction results given by black-box model in a local and global scale. From a legal perspective, the European Commission requires all artificial intelligence systems to be transparent for high risk projects (Chavanel, 2022). Meanwhile, Naser (2021) pointed out the necessities to develop explainable ML models in civil engineering from an engineering perspective.

While interpretable approaches like XAI may help engineers understand results coming from ML models, domain knowledge should also be valued. In order to understand how the field engineers rank the input parameters, we still invited a group of SNCF engineers for a survey. They were asked to grade numerically the importance of features for scour risk assessment. The specific research methodology for interpreting the machine learning model is presented in the next subsection.

6.2.2 Research methodology

Figure 6.1 illustrates the schema of the research methodology in this section. The imbalanced dataset issue observed in Table 3.2 is addressed firstly via the synthetic minority oversampling technique (SMOTE) technique. Later, two XAI approaches (SHAP and surrogate model) and engineers' expertise are employed to interpret the ML classifier, which is trained after data oversampling. In the end, the feature importance obtained from XAI approaches and engineers' expertise are compared.

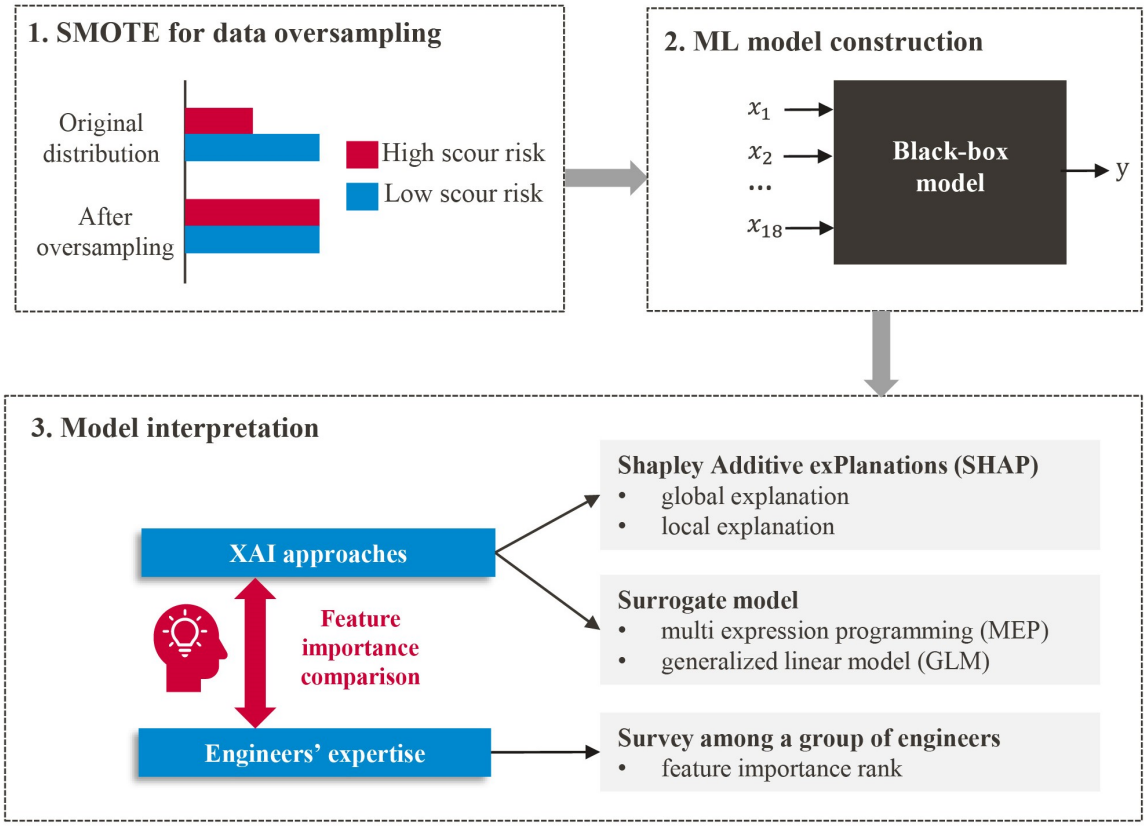


Figure 6.1: Methodology for interpreting the ML model

6.2.3 SMOTE for data oversampling

An imbalanced dataset could result in a negative effect on the performance of ML algorithms. This issue, nonetheless, exists in many real-world applications (Gosain

and Sardana, 2019). Collecting data of railway bridges requires the efforts of people in different affiliations and may sometimes be time-consuming. Consequently, it is not quite easy to obtain a balanced dataset, which means that the number of two classes is not equally distributed in the end.

In recognition of this issue, SMOTE (Chawla et al., 2002) is used in our study, which oversamples the minority class samples. Recently, diverse types of ML problems have applied this approach and proved the effectiveness (Blagus and Lusa, 2013; Naser and Kodur, 2022). In SMOTE, synthetic data is generated based on real observations. The algorithm takes firstly samples from the existing minority class. Later, new data is synthesized by calculating the distance between its two nearest neighbours. The distance is then multiplied by a random number ranging from 0 to 1 and the newly generated synthetic data is placed from one of the real observations using the calculated distance in the end.

After SMOTE, the size of bridge pier dataset is increased from 208 to 302, and the high – low scour risk classes are equally distributed (from 73% - 27% to 50% - 50%). 70% of the oversampling data is used to train the XGBoost classifier and the rest 30% is for evaluation.

Table 6.1 shows the accuracy, precision, recall and FPR of the XGBoost classifier with training and test sets respectively. Furthermore, the precision and recall scores of each class before and after SMOTE are shown in Table 6.2. It should be reminded that the ideal value for accuracy, precision and recall are equal to 1 whereas 0 for FPR.

Table 6.1: XGBoost classifier performance (data after SMOTE)

	Accuracy	**Precision**	**Recall**	**FPR**
Training set	0.99	0.98	0.99	0.02
Test set	0.95	0.95	0.93	0.04

Table 6.2: Precision and recall for each class before and after oversampling

Class		**Training set**		**Test set**	
		Precision	**Recall**	**Precision**	**Recall**
Low scour risk	Original data	0.97	0.90	0.73	0.92
	After SMOTE	0.99	0.98	0.94	0.96
High scour risk	Original data	0.96	0.99	0.98	0.92
	After SMOTE	0.98	0.99	0.95	0.93

From Table 6.1, it is observed that the XGBoost algorithm has a satisfactory performance regarding all four proposed measurements. Moreover, Table 6.2 shows that there is a significant increase regarding precision score for low scour risk classes after SMOTE. In general, with data after SMOTE, the XGBoost classifier has an encouraging performance, and the performance of ML model on minority class prediction is improved. The model interpretation is developed based on this classifier.

6.2.4 SHAP model interpretation

The feature importance values are often inconsistent by changing the criteria (Bakouregui et al., 2021), which could result in contradictory explanations. SHAP (SHapley Additive exPlanation) model, on the other hand, provides a unified measurement (Lundberg and Lee, 2017) for model interpretations. The benefits of SHAP model are

that both global and local interpretability can be achieved. This subsection presents briefly SHAP model and its interpretation results.

6.2.4.1 SHAP model introduction

SHAP (SHapley Additive exPlanation) model (Lundberg and Lee, 2017), derived from the Shapley value (Shapley, 1952) in the game theory, is an important and versatile tool in XAI. It is capable of interpreting the output of any ML model on a global and local scale.

SHAP model uses the additive feature contribution function to explain the unknown ML model. Suppose that f is the original prediction model and g is the explanation model, x' is the simplified input term that links the original input variables s through a mapping function $x = h_x(x')$. SHAP defines the explanation as:

$$f(\mathbf{x}) = g(\mathbf{x}') = \phi_0 + \sum_{j=1}^{M} \phi_j {x_j}' \tag{6.1}$$

where M is the number of simplified input variables. $\phi_j \in \mathbb{R}$ denotes feature j's contribution. ϕ_0 is the constant value when all input parameters are missing. Equation (6.1) is illustrated in Figure 6.2, where ϕ_0, ϕ_1, ϕ_2 and ϕ_3 increase the predicted result $f(x)$ while ϕ_4 decreases $f(x)$.

According to Lundberg and Lee (2017), the possible solution of equation (6.1) which guarantees the local accuracy, missingness, and consistency of the SHAP model is shown in equation (6.2):

$$\phi_j(f, \mathbf{x}) = \sum_{\mathbf{z}' \subseteq \mathbf{x}'} \frac{|\mathbf{z}|!\,(M - |\mathbf{z}'| - 1)!}{M!} \left[f_x(\mathbf{z}') - f_x(\mathbf{z}' \setminus j) \right] \tag{6.2}$$

where $|\mathbf{z}'|$ represents the number of non-zero entries in $\mathbf{z}'$ and $\mathbf{z}' \subseteq \mathbf{x}'$ $f_x(\mathbf{z}') = f(h_x(\mathbf{z}')) = E(f(\mathbf{z}) \mid \mathbf{z_s})$, and S is the set of non-zero indices in $\mathbf{z}'$. ϕ_j is the Shapely

value of feature j. More detailed information and mathematical deductions of SHAP model may refer to the work of Lundberg and Lee (2017).

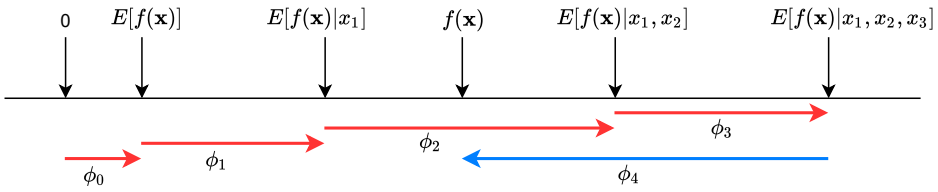


Figure 6.2: SHAP attributes regarding each feature

6.2.4.2 SHAP model interpretation results

- **Global explanation**

Figure 6.3 illustrates the impact of input features on model predictions on a global scale. The global feature importance plot, where the importance of each feature is calculated as the mean absolute SHAP value for that feature over all given samples is shown in Figure 6.3 (a). The y-axis plots the ordered feature importance rank and the x-axis stands for the mean absolute SHAP value or the average impact on model output for each feature. In general, the higher the mean absolute SHAP value is, the more important the variable is. From Figure 6.3 (a), it can be seen that the four most important features are respectively I17 (existence of local scour), C2 (slope of riverbed), I16 (existence of dislocation and deformation) and C3 (specific flood flow), while C6 (flow sinuosity), C5 (topography) and B9 (foundation type) have SHAP values equal to zero, which means they have little contribution to prediction.

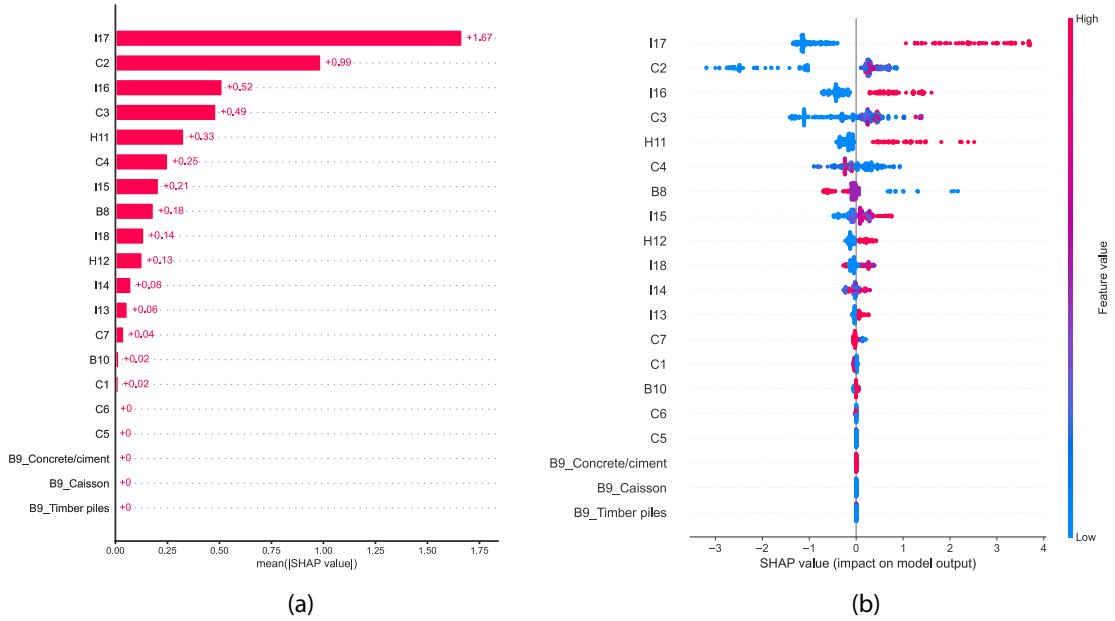


Figure 6.3: SHAP global bar plot and summary plot: (a) global bar plot; (b) summary plot.

The global bar plot in Figure 6.3 (a) only permits to see the contribution of each variable to prediction results. Nonetheless, the impact of each feature (whether it's positive or negative) to prediction is still not clear.

In this circumstance, Figure 6.3 (b) presents the summary plot. It uses an information-dense way to show the relationship between feature value (colour red means value is high while blue is low) and SHAP value (plotted in x-axis) regarding each input variable. Positive SHAP value means the positive impact (e.g., high scour risk) on prediction whereas negative value means negative effect (e.g., low scour risk). The feature value for categorical variable is presented by the encoding value shown in Figure 3.6. Each data is represented by a dot on each feature row. The "pile up" dots show the density in each row. For example, the existence of damages (features with encoding values equal to 1) like I17 (local scour), I16 (dislocation and deformation)

and H11 (scour history) will increase the scour risk (indicated by the positive SHAP value) whereas the nonexistence situations (features with encoding values equal to 0) will decrease the risk (indicated by the negative SHAP value).

For variables whose relationships between SHAP values and feature values are not clear in summary plot, Figure 6.4 illustrates the dependence scatter plot which shows the effect of a single feature on prediction.

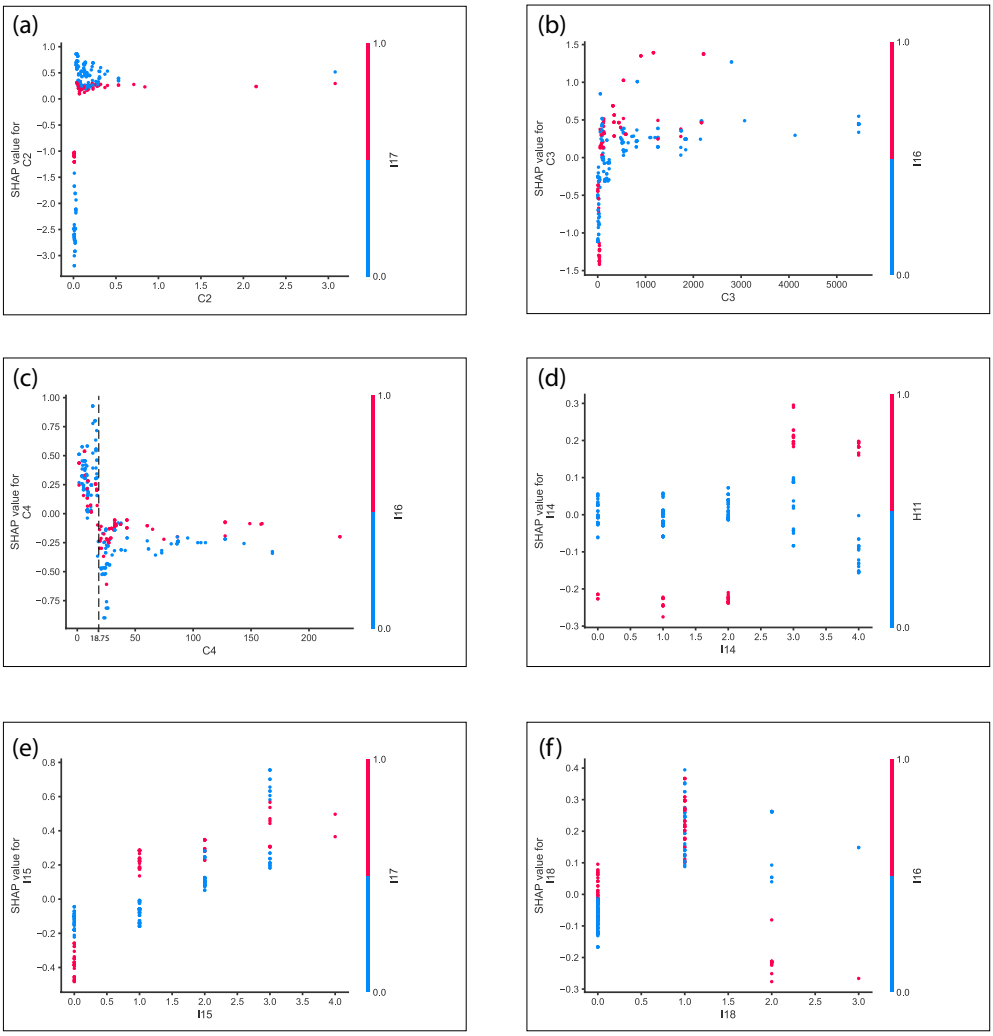


Figure 6.4: SHAP partial dependence plots: (a) C2 (slope of riverbed (%)); (b) C3 (specific flood flow (m^3/s)); (c) C4 (width of valley/ width of low flow channel); (d) I14 (channel rating); (e) I15 (riverbank rating); (f) I18 (rating of other damages)

Similarly to summary plot, each dot represents a data sample. The x-axis is the value of feature, and the y-axis is the SHAP value. Furthermore, partial dependence plot introduces a second feature (chosen automatically by default) that may have an

interaction effect with the plotted feature. Several observations are made from Figure 6.4:

1. In Figure 6.4 (c), it can be seen clearly that when C4 (ratio between width of valley and low flow channel), a parameter used to describe the hydromorphology of the region, is less than 18.75, it tends to increase the scour risk and vice versa. This finding could provide insights for engineers since the correlation between C4 and scour risk is difficult to be quantified with current knowledge.

2. The relationships between SHAP values and categorical variables I14 (channel rating) and I18 (rating of other damages) are less obvious as shown in Figure 6.4 (d) and (f).

3. Figure 6.4 (a) and Figure 6.4 (b) show that when the value of C2 (slope of riverbed) or C3 (specific flood flow) is large, it tends to increase the scour risk (SHAP value is positive), which is in accordance with the current knowledge: scour exists more often in the steep slope and/or high flow area. Furthermore, it can be observed that the feature values of C2 and C3 are significantly scattered and, consequently, the predictions as well as explanations may be less precise for areas having less value.

Figure 6.5 displays the waterfall plots for local explanations regarding data sample No.1 and No.204. The bottom of a waterfall plot shows the base value, which is considered the prediction result when input features' values are unknown. In other words, base value is the average of target variables. Then each row shows the contribution (red means positive and blue means negative) of each feature to the model output. The corresponding value of each feature is marked at the left side in grey. It should be noted that by default, the x-axis of waterfall plot is in log-odds unit. The negative value indicates the probability is less than 0.5 which means the negative class. For example, in Figure 6.5 (a), the slope of river bed (C2 = 0.01) and the nonexistence

of local scour (I17 = 0) are the two most significant factors to decrease the scour risk, while in Figure 6.5 (b) the existence of local scour (I17 = 1), the existence of dislocation or deformation (I16 = 1), slope of riverbed (C2 = 0.06), existence of scour history (H11 = 1) are the four most important aspects to push up the base value and increase the risk.

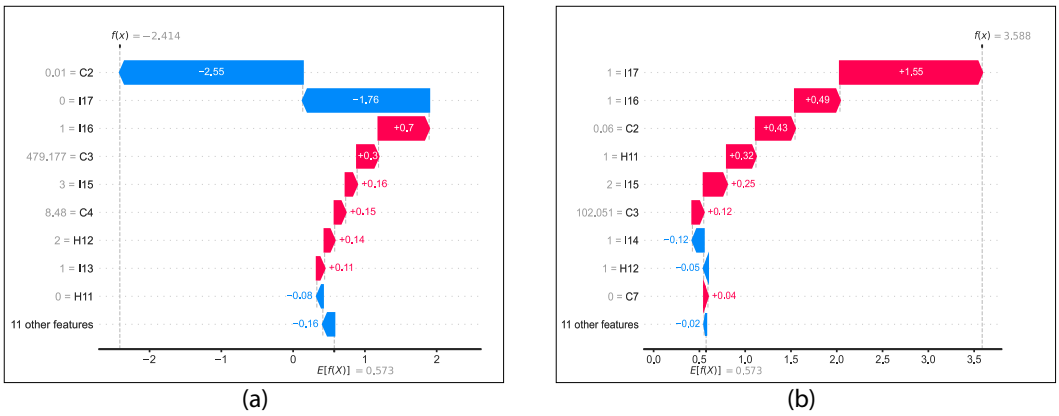


Figure 6.5: SHAP waterfall plot for local explanation: (a) data No. 1; (b) data No. 204.

To conclude, SHAP interpretation results provide insights to ameliorate current field inspections and are generally in accordance with the scientific knowledge.

6.2.5 Surrogate model interpretation

In engineering, when the prediction results are derived from a time-consuming and computationally expensive process, a surrogate model is desired to approximate the existing complex method. Compared with a black-box model, a surrogate model usually has a simple nature or the inner mechanism has already been understood by users (e.g. a linear function) (Naser, 2021). Theoretically, any model having high interpretability can be considered as a surrogate model if it can approximate the black-box model as closely as possible (Molnar, 2019). To build a surrogate model, data used to train the black-box model are input variables. The target variables,

however, come from the predictions given by the black-box model instead of original observations. With its explicit expressions, the surrogate model is considered to be able to explain the predictions once it goes through the training and test process and achieves satisfactory performance.

Multi expression programming (MEP) (Oltean and Dumitrescu, 2002) and generalized linear model (GLM) are used in our study to build two surrogate models.

MEP, as a branch of genetic programming (GP) inspired by Charles Darwin's theory of natural selection, is a type of linear-based GP for optimization. Multiple solutions (programs) are encoded in the same chromosome in MEP. It starts by generating random population of computer programs and the best solution can be generated from the chromosome by iterating the selection, crossover and mutation process until the termination condition is satisfied. The best equations generated from MEP are usually easy to be applied in practice. The expressions obtained from MEP to decide the scour risk class are shown as follows in equation (6.3) – equation (6.7):

$$A_1 = C_4 I_{17} \tag{6.3}$$

$$A_2 = C_2 + I_{16} + I_{17} - B_8 \tag{6.4}$$

$$A_3 = (B_{10} + I_{16})^{H_{11}} \tag{6.5}$$

$$A_4 = A_1 + A_2 + A_3 \tag{6.6}$$

$$\text{If } A_4 < -0.982, \text{low risk, or else high risk} \tag{6.7}$$

Risk in the end is decided by an IF function. The model achieves the following accuracy (0.93/0.9), precision (0.90/0.86), recall (0.96/0.93), and FPR (0.10/0.11), for training and test data respectively.

GLM, on the other hand, is a linear model that generalizes variables in both nu-

merical and categorical forms via a link function. Compared with a linear regression model in which both variables and outcomes are assumed to follow a Gaussian distribution, GLM allows variables and outcomes following the exponential family of distributions (e.g., Poisson distribution, binomial distribution). The link function relates the linear predictor and the mean of the distribution function. For example, the binomial distributed data may use the logistic link function.

Equation 6.8 shows the linear function using GLM to augment the interpretability of XGBoost classifier. Each input variable multiplies the computed coefficient. Considering it's a binary classification case, a logit function is served as the link function.

$$\begin{aligned} Risk_{scour} = logit\,(& -0.13 \times C_1 + 8.51 \times C_2 + 0.0007 \times C_3 - 0.027 \times C_4 + 0.375 \times C_6 \\ & - 0.74 \times C_7 - 3.68 \times B_8 - 1.27 \times B_{10} + 2.31 \times H_{11} + 1.71 \times H_{12} \\ & - 0.72 \times I_{13} + 0.54 \times I_{14} + 0.09 \times I_{15} + 2.91 \times I_{16} + 12.61 \times I_{17} + \\ & 0.18 \times I_{18}) \end{aligned} \tag{6.8}$$

The trained GLM model has the following metrics for accuracy (0.92/0.91), precision (0.94/0.93), recall (0.9/0.88) and FPR (0.06/0.06) for training and test set.

Due to the differential inherent computational process, equations from MEP use part of the input data including an exponential calculation and a logical model, while GLM formulates all input parameters in a linear form linking by a logit function.

In general, both MEP and GLM are easy to be applied in practice. But one should note that surrogate models are developed to explain the original ML model. They cannot have perfect fidelity with regard to the original model (Rudin, 2019). If the explanation completely adheres the computed results, this means the two models will be equal and there will be no need to have the original model at the beginning. In other words, the original model itself already has a transparent nature.

6.2.6 Engineers' interpretation

SHAP interpretation is based on the mathematical algorithm. It's still important to know feature importance from the engineering perspective. Due to its complexity and multidisciplinary nature, current scour risk evaluation procedures are highly empirical based. In practice, the decision-making process may be influenced by the subjectiveness of field engineers (e.g., working experience, educational level).

With the purpose to see how field engineers consider feature importance in a scour risk model and whether their opinions are unified, a survey was conducted. Among the 26 engineers who participate in the survey, 11 of them have been working in the inspection and maintenance sector for more than ten years. 10 of them have an experience of less than five years and the rest of them are between five to ten years. During this workshop, they were asked to grade numerically the influence of input parameters to scour risk from 1 (the least influential) to 4 (the most influential). It should be noted that parameters ranked by engineers are the same input parameters in ML model.

Figure 6.6 displays the average scores with standard deviations of input parameters ranked by SNCF engineers.

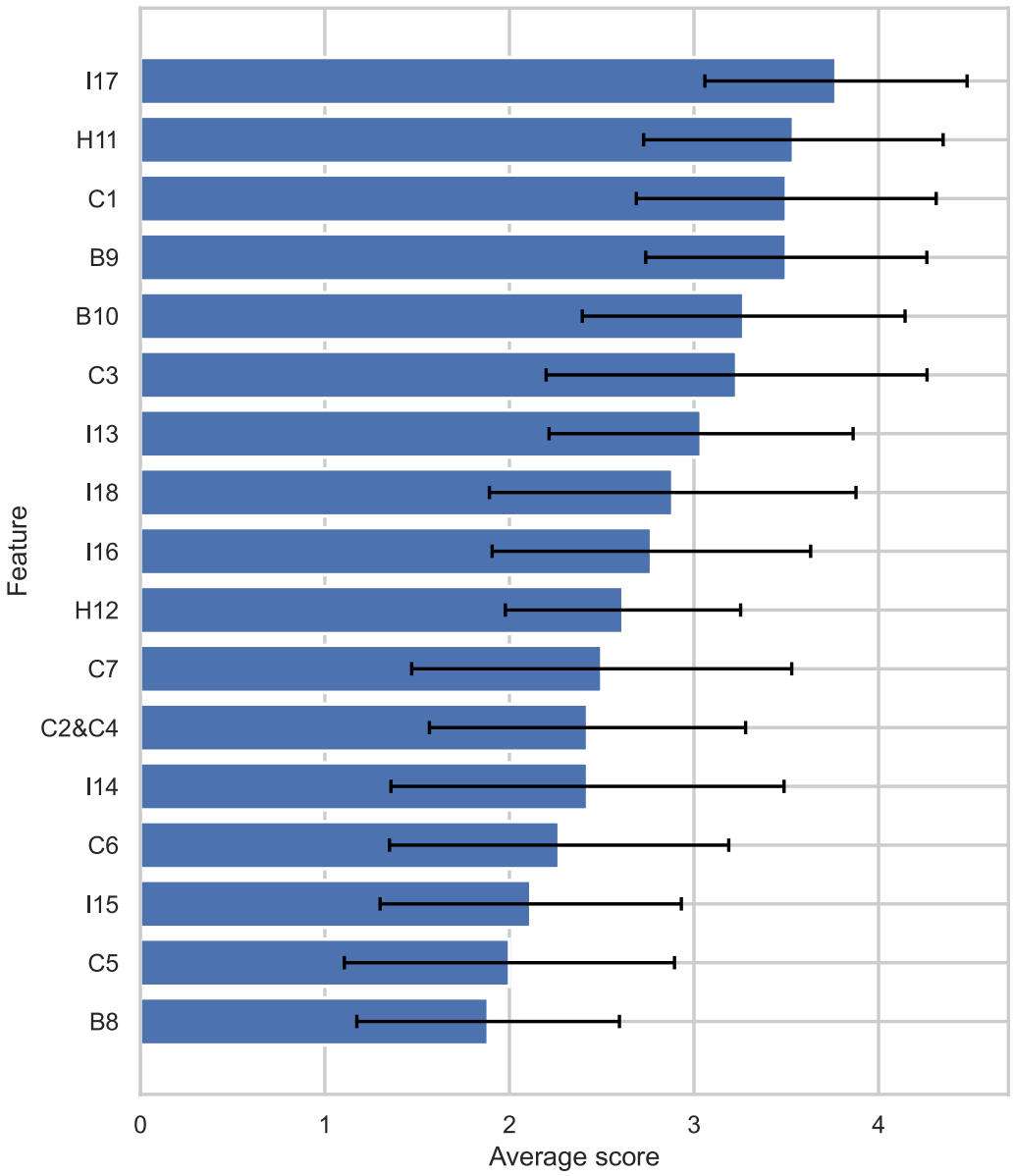


Figure 6.6: Feature average scores obtained from the survey

It is observed that almost all features have a score near or above 2 and none of them is around 1, which means variables in ML are all considered vital for decision-making. The high standard deviations existing in some features (e.g., C3, C7, I14) may be caused by working experience and working place location (contrasted climate

and geographical background from north to south in France). Furthermore, environmental factors like C2 (slope of riverbed), C3 (specific flood flow), and C4 (width of valley/width of low flow channel) are not as important as in SHAP global bar plot. Possible reasons for such differences are discussed afterwards.

6.2.7 Feature importance discussion

It is noticed in Figure 6.6 that the feature importance rank given by the engineers is not quite in consistency with the SHAP model. Therefore, Figure 6.7 compares the relative feature importance by using results from SHAP model (mean SHAP value in Figure 6.3 (a)), GLM model (coefficients in equation 6.8) and survey results (average score in Figure 6.6) with a 0 -1 normalization. MEP approach is not compared due to its exponential functions.

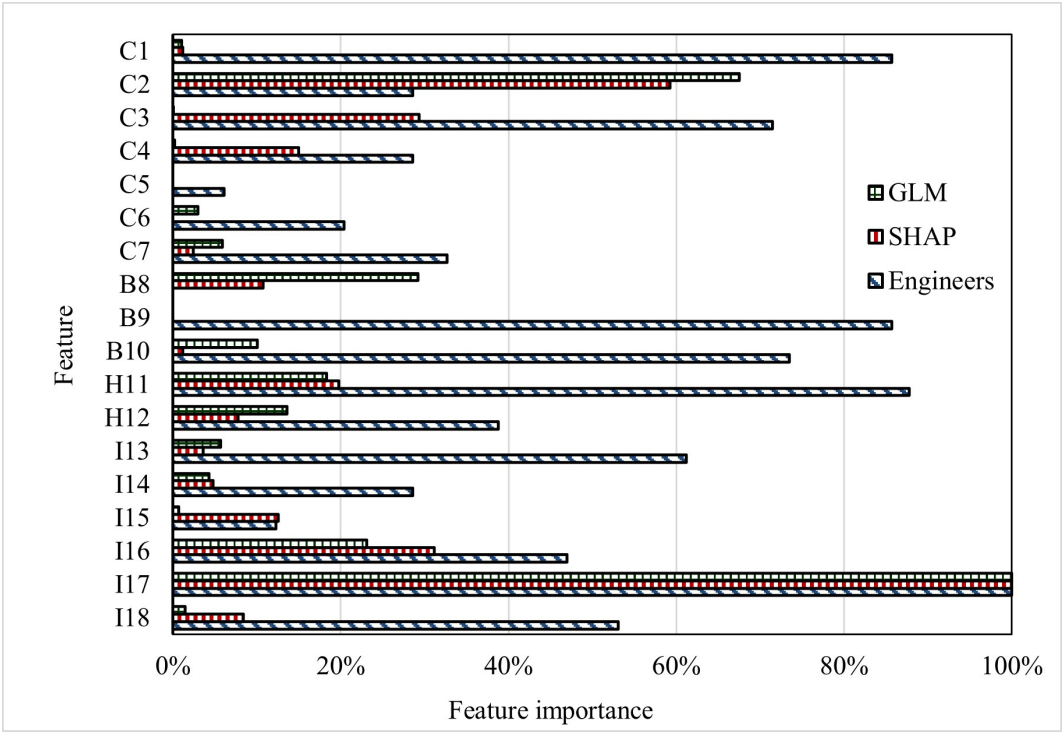


Figure 6.7: Feature importance comparison.

Several observations are made from Figure 6.7 :

1. I17 (existence of local scour) is considered as the most important feature in all three methods.

2. Some features have little or zero contribution in SHAP and GLM methods while the engineers give a rather high score, such as C1 (flow type), C5 (topography) and B9 (foundation type). It's noteworthy that these are features mentioned in commonly used scour risk evaluation process.

3. By contrast, C2 (slope of riverbed) and C4 (width of valley/width of low flow channel) have a fairly high score in SHAP but not with the engineers.

To conclude, for both XAI models and engineers, the existence of local scour (I17) is considered as the most important feature. It seems logical since estimating the local scour depth under future flood scenarios has always been a key step in most practical guidelines (e.g., Cerema, 2019; HR Wallingford, 1992). The risk could be already high if a scour hole (local scour) around bridge foundation is detected. Besides, the two XAI approaches are generally consistent with each other.

The reasons for differences between engineers' expertise and XAI approaches are concluded as follows:

1. An engineer does the risk analysis in an asset-specific view and makes judgements based on physical mechanisms. However, the ML model makes predictions by learning from data. It is the statistical relationship and algorithms that "teach" the model. In this case, the important variables for engineers may be already reflected or covered by other variables in XAI interpretations due to the different natures of methodologies. For example, flow type (C1) and topography (C5) could be reflected by slope of riverbed (C2) and hydromorphology of the region (C4). Cracks (information covered by variable I18) on bridge piers could

reflect the non-sufficient bearing capacity of bridge foundation represented by foundation type (B9).

2. The role of ML model in the whole inspection-maintenance process is to help engineers for decision-making, while the tasks for engineers are more complex than this. They still need to program the corresponding maintenance work once the asset is considered at high risk. Therefore, a comprehensive understanding shall be necessary for engineers, which includes factors such as C1 (flow type), B9 (foundation type), B10 (existence of foundation scour countermeasures).

3. The interviewees in this survey (SNCF engineers), who mainly do inspections, are not specialists in geotechnics or hydrology. Due to the staff shortage but a huge number of rail assets to be inspected, they make decisions mainly based on the damages observed in the field (e.g., scour hole, crack, material degradation) and sometimes ignore the surrounding environment. If engineers participating in the survey are specialized in hydrology, environmental factors like C2 (slope of riverbed) and C4 (width of valley/width of low flow channel) and bridge pier shape (B8) could be more important in this case.

4. In the end, although two popular XAI models were employed to explain the black-box model in this study, Tocchetti and Brambilla (2022) pointed out that state-of-the-art explainability may not be enough to guarantee the full understandability of explanations from a human perspective. Evaluating the effectiveness of explanations could require researchers from IT-related fields, psychology and philosophy, which is considered beyond the scope of this study.

In comparison with engineers' expertise, XAI approaches consider variables for describing the characteristics of watercourse (C2: slope of riverbed; C3: specific flood flow) and hydromorphology of the region (C4: width of valley/width of low flow channel) are more important than bridge characteristics (e.g., foundation type). This

is mainly because the resilience of historical bridges to natural hazards is not as good as new constructions. The same flood event could cause more severe damages to historical infrastructures. In other words, compared with structural factors, the surrounding environment could pose an even greater threat to the stability of historical bridges since all of them have already been in a relatively vulnerable state.

Interpreting the ML model via XAI and comparing the results with engineers' expertise contribute to making the ML model trustworthy and improving the current inspection process. The proposed method is capable of screening high scour risk structures effectively and helping engineers understand how the prediction is made at the same time. Furthermore, the comparison results emphasize the importance of surrounding environmental factors in scour risk assessment, which are currently often neglected by engineers.

In the end, there hasn't existed a standard to criticize whether XAI or engineers' expertise is closer to the true mechanism. The proposed XGBoost model with explanations and the survey to rank input parameters are now built in a web application at SNCF. The updated database and more engineers' participation in the survey could help better understand the phenomenon in the future.

6.3 Model implementation at SNCF

This section introduces how to implement the proposed ML classifier in practice. The XGBoost classifier we've built and the SHAP model interpretation have been included in a web application [1], which could be directly used by the engineers without knowing coding.

[1] This thesis proposed the proof of concept of the application. Development work was supported by the SNCF PLATIPUS project.

6.3.1 Work flow

Figure 6.8 depicts schematically the workflow of the proposed model in practice at SNCF. Data collected by engineers through inspection is firstly recorded in an SNCF internal and digitalized platform named PIGC (Patrimoine Informatisé du Génie Civil), which is designated for managing rail infrastructure. Information in this platform regarding the bridge characteristics, surrounding environment as well as the observations from inspections is used as data to train the ML model. After training and testing, the scour risk provided by the ML model is seen as a complement to the engineer's assessment. Corresponding scour countermeasure work could be scheduled hereafter.

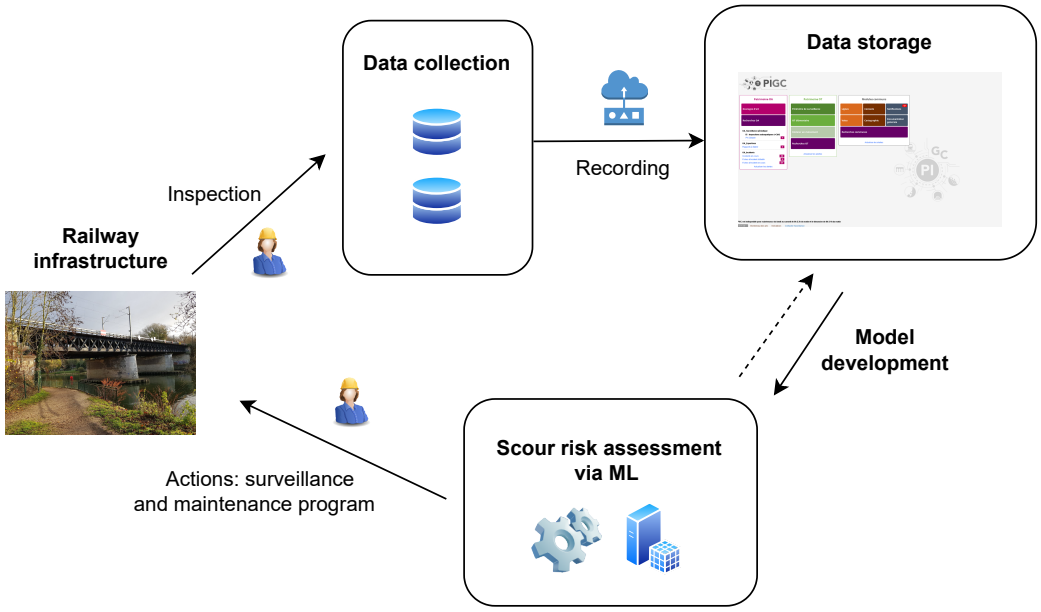


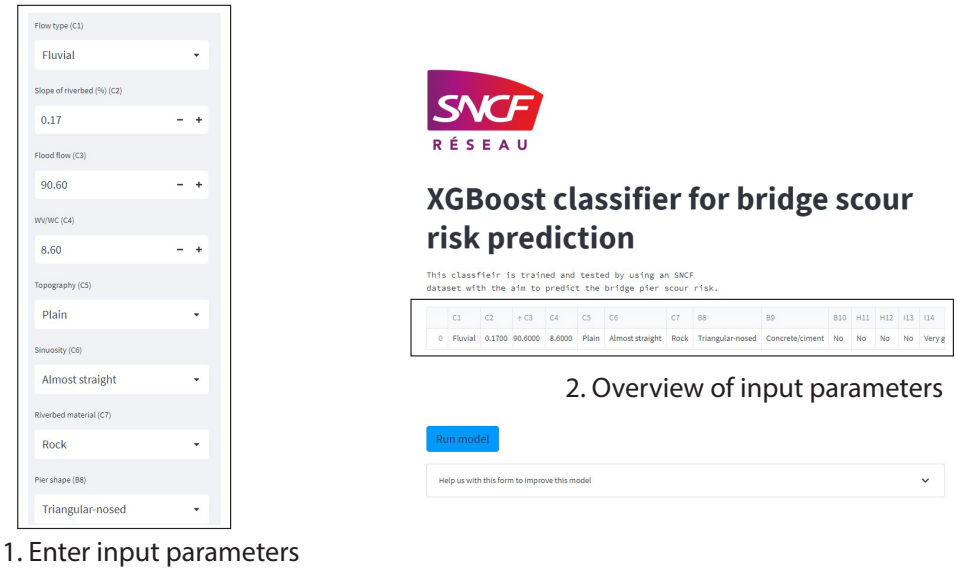
Figure 6.8: Overview of the machine learning based PdM system

6.3.2 Web application

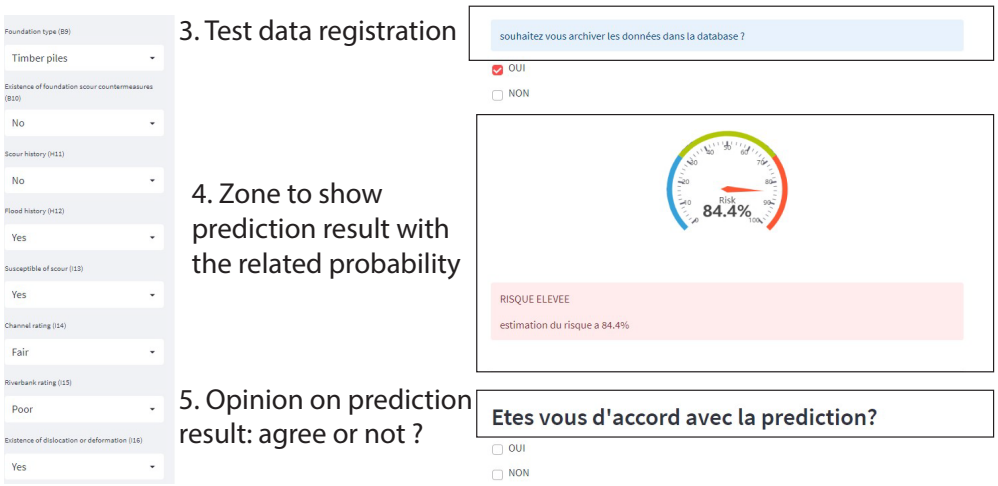
Building the proposed ML model requires solid knowledge not only in the domain of civil engineering but also coding. For the users, namely the inspectors and engineers,

it's very rare for them to have the two competences at the same time. Therefore, the proposed ML classifier has been employed in a web application, as shown in Figure 6.9. It could help the inspectors and engineers directly use the model without having a background in coding.

After inspections, the engineer can enter the input parameters based on their field observations. An overview for confirming the entered parameters is shown in the web application (see Figure 6.9 (a)). Before knowing the result, users are asked firstly if they are willing to register the test data in the cloud. For developers, the registered data could help enrich the database and update the ML model regularly with the aim of maintaining the system. The prediction result with related probability is shown afterwards. In order to improve the ML model, a question is asked to know whether the users agree with the prediction results or not (see Figure 6.9 (b)). SHAP local interpretation has been implemented. It helps understand how the prediction is made by the ML classifier. In the end, the survey conducted among the engineers to rank feature importance has been included in this web application as well. It allows collecting more data from the engineers (see Figure 6.9 (c)).

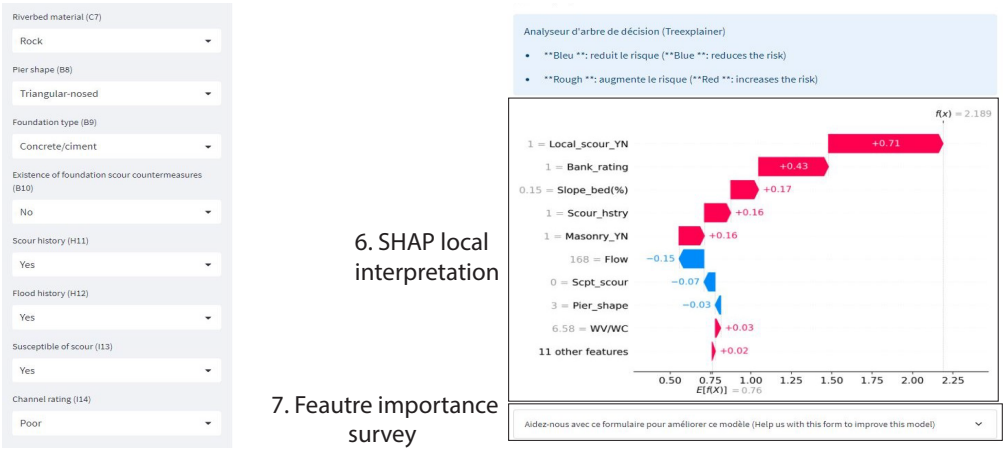


(a) Zone for entering input parameters



(b) Prediction result and collecting feedback from the users

Figure 6.9: Proposed ML model in a web application (Copyright SNCF)



(c) Implementation of SHAP local interpretation

Figure 6.9: Proposed ML model in a web application (Copyright SNCF)

The web application we have built not only allows engineers who don't know coding can use it easily, but also helps the maintenance, data collection and future improvement of the model.

6.4 Conclusions

The XGBoost classifier that we have built in Chapter 4 could help field engineers evaluate scour risk rapidly yet precisely. Although satisfactory results have been obtained, it is difficult to apply such model in practice due to the opaque nature of ML algorithms, from both engineering and legal perspectives.

In this circumstance, this chapter firstly proposed interpretable approaches to make the black-box ML model transparent by using XAI and engineers' expertise. Data after oversampling was at first used to build the ML classifier. Later, XAI approaches (SHAP and surrogate models) and a survey among SNCF engineers were conducted to interpret the ML model. SHAP proposed global and local explanations for model interpretations. Surrogate models namely MEP and GLM approximated the original

model and showed satisfactory results by generalizing functions easy to be used in the field. The engineers' survey results indicate that the environmental factors are not as important as in SHAP global plot. In the end, feature importance obtained from SHAP, GLM and engineers' survey were compared.

The comparison results show that existence of local scour (variable I16) around bridge foundation is considered as the most important feature in XAI approaches and engineers' expertise, which is in line with current knowledge: one of the most obvious criteria to assess scour risk is to know whether there has already been a scour hole near bridge foundation. Besides this common point, it is observed that several environmental factors neglected by engineers have a rather high ranking in XAI explanations. Generally speaking, the differences between XAI approaches and engineers' expertise are caused by different natures of methodologies, scope of work, and bias existing in the engineers.

In this chapter, the XAI interpretations make ML model trustworthy through explaining how the predictions are made. The survey among the engineers reveals the fact that some hydrological parameters are not prioritised by the engineers during scour risk evaluation. In the end, comparing XAI results with engineers' expertise helps significantly in improving the maintenance process at SNCF. It highlights the importance of surrounding environmental factors (watercourse, riverbed, hydromorphology). Compared with new constructions, historical bridges are in a relatively vulnerable state and the surrounding environment could pose a greater threat to their stability.

In the end, how to implement the proposed machine learning model in practice is presented. A web application is built using the research outcome of the thesis. It allows engineers using and understanding the model easily, even without knowing coding. Moreover, this web application serves as the bridge of communication between model developers and engineers. The feedback from users will help maintain and

improve the model. Future work should continue enlarging the size of data and inviting more railway engineers to participate in the survey.

Chapter 7

Conclusions

7.1 General conclusions

Due to the increase of precipitation brought by climate change, flood events are very likely to happen more often in the future. Flooding is actually the direct cause of scour induced infrastructure failure. Huge amount of money and resources are deployed each year to reinforce or maintain the infrastructure subject to scour. SNCF is currently seeking a more appropriate approach for evaluating the scour risk.

To address this issue, machine learning based solutions are proposed in this thesis. Two datasets are established by using information from inspection reports and open source platforms. In order to determine the input parameters for the machine learning models, feature selection work is conducted from both engineering and statistical perspectives.

Later, very popular and commonly used algorithms are employed. The performance of machine learning classifiers are examined by using measurements calculated from the confusion matrix. Results have shown that for the bridge pier dataset, XGBoost

and RF models have the most promising results. Regarding the Abutment&Wall dataset, due to the lack of data, an overfitting is observed at first. After eliminating several input parameters, it's still the XGBoost algorithm who has the best performance. But generally speaking, the prediction accuracy and other measurements for Abutment&Wall dataset are not as good as in pier dataset.

Next, in order to compare the models, more complex investigations are conducted regarding the RF and XGBoost classifiers trained for the bridge pier dataset because they achieve satisfying results. It is observed that in most cases, both models can perform in a robust fashion. XGBoost classifier is more sensible for data uncertainty but it is the one that has the most common results (95%) as the senior engineer. When tested by the Japanese cases, RF model is the one that has the most cases in common (73.3%) with the Japanese guideline. This is because RF has a relatively high recall score and the Japanese bridges are in rather vulnerable states when judging by the French standard. Moreover, when applying ARPSA to French railway bridges, it is found ARPSA could easily overestimate the scour risk. That is to say, although ARPSA is built for French road bridges who could possibly share very similar geographical environment as the railway ones, it can not be directly used. One of the reasons is that the threshold for vulnerability level classification is empirical based and it's not perfectly adaptable in the rail sector.

Then, in order to understand the predictions from machine learning model, the XGBoost model for bridge pier dataset is interpreted by SHAP model, surrogate model, and engineering expertise. The interpretation results are compared afterwards. The differences between XAI approaches and engineer's expertise are caused by the different natures of methodologies, scope of work, and bias existing in the engineers. This comparison highlights the importance of surrounding environmental factors such as hydromorphology, riverbed, watercourse, which are currently often less prioritised by the engineers or inspectors.

In the end, to facilitate the use of engineers and inspectors, the XGBoost classifier with XAI explanations is built in a web application. Findings presented in this work could help significantly evaluate the scour risk by benefiting from the novel AI and machine learning technology. It could provide valuable guidance for improving the current inspection process and insights for other countries who want to develop their own practical guideline.

7.2 Perspectives

Several recommendations are made for the future work.

1. This study can be seen as a first try to use machine learning for scour risk evaluation. A binary classifier is built with limited number of data. However, it is still difficult to prioritize maintenance work considering the number of rail assets. A multiclass model for risk segmentation shall be more practical in practice.

2. The size of the two datasets is relatively small considering the number of assets in the rail network. Therefore, future work should continue enlarging the size of data, especially for the Abutment&Wall dataset. More variables could possibly be included if more data is collected in the future. The ideal algorithm for making the prediction could possibly be changed as well.

3. The proposed model is capable of identifying scour vulnerable bridges from a technical perspective. However, the maintenance activities should be planned long time in advance. With constraint budget and the shortage of stuff, the asset manager should get to know the cost (direct and indirect) associated with. The cost of making a bridge less vulnerable to scour is smaller compared to the total cost of failure. Future analysis should include corresponding cost-benefit studies and fortunately, some studies have already mentioned this point (Liu

et al., 2020; Wright et al., 2012).

4. In the upcoming year, the changing climate will undoubtedly pose a greater threat to the safety of rail infrastructure and it will especially be amplified in long service life bridges, which are built long time ago and the impact of climate change on the intensity of flood actions are not taken into account in design phase. Therefore, it shall be important to incorporate the effects of climate change and project the bridge scour risk under future climate scenarios for the safety of transport network.

Bibliography

Abedi, M., Naser, M.Z., 2021. RAI: Rapid, Autonomous and Intelligent machine learning approach to identify fire-vulnerable bridges. Applied Soft Computing 113, 107896. doi:10.1016/j.asoc.2021.107896.

Adoko, A.C., Jiao, Y.Y., Wu, L., Wang, H., Wang, Z.H., 2013. Predicting tunnel convergence using multivariate adaptive regression spline and artificial neural network. Tunnelling and Underground Space Technology 38, 368–376. doi:https://doi.org/10.1016/j.tust.2013.07.023.

Alipour, M., Harris, D.K., Barnes, L.E., Ozbulut, O.E., Carroll, J., 2017. Load-Capacity Rating of Bridge Populations through Machine Learning: Application of Decision Trees and Random Forests. Journal of Bridge Engineering 22, 04017076. doi:10.1061/(asce)be.1943-5592.0001103.

Amini, A., Melville, B.W., Ali, T.M., Ghazali, A.H., 2012. Clear-water local scour around pile groups in shallow-water flow. Journal of Hydraulic Engineering 138, 177–185. doi:10.1061/(ASCE)HY.1943-7900.0000488.

Annandale, G.W., 2006. Scour technology : mechanics and engineering practice. McGraw-Hill, New York.

Argyroudis, S.A., Mitoulis, S.A., 2021. Vulnerability of bridges to individual and multiple hazards- floods and earthquakes. Reliability Engineering and Sys-

tem Safety 210, 107564. URL: https://doi.org/10.1016/j.ress.2021.107564, doi:10.1016/j.ress.2021.107564.

Armaghani, D.J., Faradonbeh, R.S., Rezaei, H., Rashid, A.S.A., Amnieh, H.B., 2018. Settlement prediction of the rock-socketed piles through a new technique based on gene expression programming. Neural Computing and Applications 29, 1115–1125. doi:10.1007/s00521-016-2618-8.

Arneson, L.A., Zevenbergen, L., Lagasse, P., Clopper, P., 2012. Evaluating Scour at Bridges. HEC-18. Fifth Edition, Hydraulic Engineering Circular No. 18. Publication. Technical Report 18.

Aurélien Géron, 2019. Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. URL: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.

Bakouregui, A.S., Mohamed, H.M., Yahia, A., Benmokrane, B., 2021. Explainable extreme gradient boosting tree-based prediction of load-carrying capacity of FRP-RC columns. Engineering Structures 245, 112836. doi:10.1016/j.engstruct.2021.112836.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion 58, 82–115. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

Bateni, S.M., Jeng, D.S., Melville, B.W., 2007. Bayesian neural networks for prediction of equilibrium and time-dependent scour depth around bridge piers. Advances in Engineering Software 38, 102–111. doi:https://doi.org/10.1016/j.advengsoft.2006.08.004.

Bejarbaneh, B.Y., Bejarbaneh, E.Y., Amin, M.F.M., Fahimifar, A., Jahed Armaghani, D., Majid, M.Z.A., 2018. Intelligent modelling of sandstone deformation behaviour using fuzzy logic and neural network systems. Bulletin of Engineering Geology and the Environment 77, 345–361. doi:10.1007/s10064-016-0983-2.

Benn, J., 2013. Railway bridge failure during flooding in the uk and ireland. Proceedings of the Institution of Civil Engineers - Forensic Engineering 166, 163–170. doi:10.1680/feng.2013.166.4.163.

Bento, A.M., Gomes, A., Viseu, T., Couto, L., Pêgo, J.P., 2020. Risk-based methodology for scour analysis at bridge foundations. Engineering Structures 223, 111115. doi:10.1016/j.engstruct.2020.111115.

Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013 14:1 14, 1–16. doi:10.1186/1471-2105-14-106.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Taylor Francis. URL: https://books.google.fr/books?id=JwQx-WOmSyQC.

British Highways Agency, 2012. The Assessment of Scour and Other Hydraulic Actions at Highway Structures (BD97/12) 3, 54. URL: http://www.standardsforhighways.co.uk/ha/standards/dmrb/vol3/section4/%0Abd9712.pdf.

Cattan, J., Mohammadi, J., 2002. Analysis of Bridge Condition Rating Data Using Neural Networks. Computer-Aided Civil and Infrastructure Engineering 12, 419–429. doi:10.1111/0885-9507.00074.

Cerema, 2019. Analyse de risque des ponts en site affouillable (in French).

Chavanel, C., 2022. L'IA pour le secteur ferroviaire européen - Etat des lieux et perspectives (in French). Reuvue Générale des chemins de fer , 89–104.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research 16, 321–357. doi:10.1613/jair.953, arXiv:1106.1813.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. doi:10.1145/2939672.2939785, arXiv:1603.02754.

Cheng, M.Y., Cao, M.T., 2015. Hybrid intelligent inference model for enhancing prediction accuracy of scour depth around bridge piers. Structure and Infrastructure Engineering 11, 1178–1189. doi:10.1080/15732479.2014.939089.

Chollet, F., et al., 2015. Keras. URL: https://github.com/fchollet/keras.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297. doi:10.1007/bf00994018.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press. doi:10.1017/cbo9780511801389.

Dahigamuwa, T., Yu, Q., Gunaratne, M., Sheng, Y., Martinez-Frias, J., 2016. Feasibility Study of Land Cover Classification Based on Normalized Difference Vegetation Index for Landslide Risk Assessment. Geosciences 2016, Vol. 6, Page 45 6, 45. doi:10.3390/GEOSCIENCES6040045.

Deng, L., Cai, C.S., 2010. Bridge Scour: Prediction, Modeling, Monitoring, and Countermeasures — Review. Practice Periodical on Structural Design and Construction , 125–134.

Díez, J., Khalifa, K., Leuridan, B., Díez, J., Khalifa, K., Leuridan, B., 2013. Gen-

eral theories of explanation: buyer beware. Synthese 190, 379–396. doi:10.1007/s11229-011-0020-8.

Dikanski, H., Imam, B., Hagen-Zanker, A., 2018. Effects of uncertain asset stock data on the assessment of climate change risks: A case study of bridge scour in the UK. Structural Safety 71, 1–12. doi:10.1016/j.strusafe.2017.10.008.

Dong, W., Huang, Y., Lehane, B., Ma, G., 2020. Xgboost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. Automation in Construction 114, 103155. doi:https://doi.org/10.1016/j.autcon.2020.103155.

Durand, E., Davi, D., Delgado, J.L., 2019. Arosa: A new french guideline for scour at bridges risk-based analysis. Proceedings of the 9th International Conference on Scour and Erosion, ICSE 2018 , 713–720doi:10.1201/9780429020940-35.

Economist, T., 2021. Climate change made north-west Europe's lethal flood more likely. URL: https://www.economist.com/europe/2021/08/23/climate-change-made-north-west-europes-lethal-flood-more-likely.

Ekuje, F.T., 2018. Bridge scour - climate change effect and modelling uncertainties. Ph.D. thesis. University of Surrey.

Elbaz, K., Shen, S.L., Zhou, A., Yuan, D.J., Xu, Y.S., 2019. Optimization of epb shield performance with adaptive neuro-fuzzy inference system and genetic algorithm. Applied Sciences 9. doi:10.3390/app9040780.

Elhag, T.M.S., Wang, Y.M., 2007. Risk Assessment for Bridge Maintenance Projects: Neural Networks versus Regression Techniques. Journal of Computing in Civil Engineering 21, 402–409. doi:10.1061/(asce)0887-3801(2007)21:6(402).

Emmert-Streib, F., Yli-Harja, O., Dehmer, M., 2020. Explainable Artificial Intelligence and Machine Learning: A reality rooted perspective URL: https://arxiv.org/abs/2001.09464v1, arXiv:2001.09464.

Fawcett, T., 2006. An introduction to roc analysis. Pattern Recognition Letters 27, 861–874. doi:https://doi.org/10.1016/j.patrec.2005.10.010. rOC Analysis in Pattern Recognition.

Few, R., 2003. Flooding, vulnerability and coping strategies: local responses to a global threat. Progress in Development Studies 3, 43–58. doi:10.1191/1464993403ps049ra.

Foti, S., Sabia, D., 2011. Influence of foundation scour on the dynamic response of an existing bridge. Journal of Bridge Engineering 16, 295–304. doi:10.1061/(ASCE)BE.1943-5592.0000146.

Franklin, J., 2010. Mapping Species Distributions: Spatial Inference and Prediction. Ecology, Biodiversity and Conservation, Cambridge University Press. doi:10.1017/CBO9780511810602.

Froehlich, D.C., 1988. Analysis of onsite measurements of scour at piers, ASCE, New York. pp. 534–539. URL: http://pubs.er.usgs.gov/publication/70014448.

Ghasemi, E., Gholizadeh, H., 2019. Development of Two Empirical Correlations for Tunnel Squeezing Prediction Using Binary Logistic Regression and Linear Discriminant Analysis. Geotechnical and Geological Engineering 37, 3435–3446. URL: https://doi.org/10.1007/s10706-018-00758-0, doi:10.1007/s10706-018-00758-0.

Ghazvinei, P.T., Ariffin, J., Abdullah, J., Mohamed, T.A., 2014. Comparative Analysis between Observed and Predicted Contraction Scour at Bridges Abutments. Research Journal of Applied Sciences, Engineering and Technology 8, 452–459. doi:10.19026/RJASET.8.993.

Ghorbani, B., Sadrossadat, E., Bolouri Bazaz, J., Rahimzadeh Oskooei, P., 2018. Numerical ANFIS-Based Formulation for Prediction of the Ultimate Axial Load

Bearing Capacity of Piles Through CPT Data. Geotechnical and Geological Engineering 36, 2057–2076. doi:10.1007/s10706-018-0445-7.

Goerlandt, F., Montewka, J., Kuzmin, V., Kujala, P., 2015. A risk-informed ship collision alert system: Framework and application. Safety Science 77, 182–204. doi:10.1016/J.SSCI.2015.03.015.

Goh, A.T., 1994. Seismic Liquefaction Potential Assessed by Neural Networks. Journal of Geotechnical Engineering 120, 1467–1480. doi:10.1061/(ASCE)0733-9410(1994)120:9(1467).

Goh, A.T., Goh, S.H., 2007. Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. Computers and Geotechnics 34, 410–421. doi:10.1016/j.compgeo.2007.06.001.

Gosain, A., Sardana, S., 2019. Farthest SMOTE: A Modified SMOTE Approach BT - Computational Intelligence in Data Mining, Springer Singapore. pp. 309–320.

Gunning, D., 2017. Explainable artificial intelligence (XAI), Defense Advanced Research Projects Agency (DARPA). Technical Report. Defense Ad- vanced Research Projects Agency (DARPA). URL: https://scholar.google.com/scholar?hl=en&as_sdt=0,5&cluster=3705946364541381453.

Hager, W.H., Unger, J., 2010. Bridge Pier Scour under Flood Waves. Journal of Hydraulic Engineering 136, 842–847. doi:10.1061/(asce)hy.1943-7900.0000281.

Harandizadeh, H., Toufigh, M.M., Toufigh, V., 2019. Application of improved ANFIS approaches to estimate bearing capacity of piles. Soft Computing 23, 9537–9549. doi:10.1007/s00500-018-3517-y.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H.,

Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 2020 585:7825 585, 357–362. doi:10.1038/s41586-020-2649-2, arXiv:2006.10256.

Hong, J.H., Goyal, M.K., Chiew, Y.M., Chua, L.H., 2012. Predicting time-dependent pier scour depth with support vector regression. Journal of Hydrology 468-469, 241–248. URL: http://dx.doi.org/10.1016/j.jhydrol.2012.08.038, doi:10.1016/j.jhydrol.2012.08.038.

Hosseini, R., Fazloula, R., Saneie, M., Amini, A., 2018. Bagged neural network for estimating the scour depth around pile groups. International Journal of River Basin Management 16, 401–412. doi:10.1080/15715124.2017.1372449.

HR Wallingford, 1992. Hydraulic aspects of bridges: assessment of the risk of scour (EX 2502) URL: http://eprints.hrwallingford.com/id/eprint/315.

Hu, X., Shi, L., Lin, G., Lin, L., 2021. Comparison of physical-based, data-driven and hybrid modeling approaches for evapotranspiration estimation. Journal of Hydrology 601, 126592. URL: https://www.sciencedirect.com/science/article/pii/S0022169421006405, doi:https://doi.org/10.1016/j.jhydrol.2021.126592.

IPCC, 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Technical Report. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Kellermann, P., Schönberger, C., Thieken, A.H., 2016. Large-scale application of the flood damage model RAilway Infrastructure Loss (RAIL). Natural Hazards and Earth System Sciences 16, 2357–2371. doi:10.5194/nhess-16-2357-2016.

Kiureghian, A.D., Ditlevsen, O., 2009. Aleatory or epistemic? does it matter? Struc-

tural Safety 31, 105–112. doi:https://doi.org/10.1016/j.strusafe.2008.06.020.

Kordjazi, A., Pooya Nejad, F., Jaksa, M.B., 2014. Prediction of ultimate axial load-carrying capacity of piles using a support vector machine based on CPT data. Computers and Geotechnics 55, 91–102. doi:10.1016/j.compgeo.2013.08.001.

Lamb, R., Garside, P., Pant, R., Hall, J.W., 2019. A Probabilistic Model of the Economic Risk to Britain's Railway Network from Bridge Scour During Floods. Risk Analysis 39, 2457–2478. doi:https://doi.org/10.1111/risa.13370.

Laursen, E.M., 1963. An analysis of relief bridge scour. Journal of the Hydraulics Division 89, 93–118. doi:10.1061/JYCEAJ.0000896.

Laursen, E. M., Toch, A., 1956. Scour around bridge piers and abutments.

Lee, T.L., Jeng, D.S., Zhang, G.H., Hong, J.H., 2007. Neural Network Modeling for Estimation of Scour Depth Around Bridge Piers. Journal of Hydrodynamics 19, 378–386. doi:10.1016/S1001-6058(07)60073-0.

Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 247, 124–136. doi:https://doi.org/10.1016/j.ejor.2015.05.030.

Li, L., Wang, J., Leung, H., Jiang, C., 2010. Assessment of catastrophic risk using bayesian network constructed from domain knowledge and spatial data. Risk Analysis 30, 1157–1175. doi:10.1111/j.1539-6924.2010.01429.x.

Li, Y., Chen, G., Tang, C., Zhou, G., Zheng, L., 2012. Rainfall and earthquake-induced landslide susceptibility assessment using GIS and Artificial Neural Network. Natural Hazards and Earth System Sciences 12, 2719–2729. URL: https://nhess.copernicus.org/articles/12/2719/2012/, doi:10.5194/nhess-12-2719-2012.

Link, O., Castillo, C., Pizarro, A., Rojas, A., Ettmer, B., Escauriaza, C., Manfreda, S., 2017. A model of bridge pier scour during flood waves. Journal of Hydraulic Research 55, 310–323. doi:10.1080/00221686.2016.1252802.

Liu, B.y., Ye, L.y., Xiao, M.l., Miao, S., 2006. Artificial Neural Network Methodology for Soil Liquefaction Evaluation Using CPT Values, in: Proceedings of the 2006 International Conference on Intelligent Computing - Volume Part I, Springer-Verlag, Berlin, Heidelberg. pp. 329–336. doi:10.1007/11816157_36.

Liu, L., Yang, D.Y., Frangopol, D.M., 2020. Network-Level Risk-Based Framework for Optimal Bridge Adaptation Management Considering Scour and Climate Change. Journal of Infrastructure Systems 26, 1–15. doi:10.1061/(asce)is.1943-555x.0000516.

Liu, Z., Gilbert, G., Cepeda, J.M., Lysdahl, A.O.K., Piciullo, L., Hefre, H., Lacasse, S., 2021. Modelling of shallow landslides with machine learning algorithms. Geoscience Frontiers 12, 385–393. doi:https://doi.org/10.1016/j.gsf.2020.04.014.

Lu, J.Y., Hong, J.H., Su, C.C., Wang, C.Y., Lai, J.S., 2008. Field measurements and simulation of bridge scour depth variations during floods. Journal of Hydraulic Engineering 134, 810–821. doi:10.1061/(ASCE)0733-9429(2008)134:6(810).

Lundberg, S., Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 2017-December, 4766–4775. URL: https://arxiv.org/abs/1705.07874v2, arXiv:1705.07874.

López, G., Teixeira, L., Ortega-Sánchez, M., Simarro, G., 2014. Estimating final scour depth under clear-water flood waves. Journal of Hydraulic Engineering 140, 328–332. doi:10.1061/(ASCE)HY.1943-7900.0000804.

Ma, G., Qin, C., Hwang, H.J., Zhou, Z., 2023. Data-driven models for predicting tensile load capacity and failure mode of grouted splice sleeve connection. En-

gineering Structures 289, 116236. doi:https://doi.org/10.1016/j.engstruct.2023.116236.

Mahdevari, S., Shahriar, K., Yagiz, S., Akbarpour Shirazi, M., 2014. A support vector regression model for predicting tunnel boring machine penetration rates. International Journal of Rock Mechanics and Mining Sciences 72, 214–229. doi:https://doi.org/10.1016/j.ijrmms.2014.09.012.

Mahdevari, S., Shirzad Haghighat, H., Torabi, S.R., 2013. A dynamically approach based on svm algorithm for prediction of tunnel convergence during excavation. Tunnelling and Underground Space Technology 38, 59–68. doi:https://doi.org/10.1016/j.tust.2013.05.002.

Mangalathu, S., Hwang, S.H., Jeon, J.S., 2020. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Engineering Structures 219, 110927. doi:10.1016/j.engstruct.2020.110927.

Mbarak, W.K., Cinicioglu, E.N., Cinicioglu, O., 2020. SPT based determination of undrained shear strength: Regression models and machine learning. Frontiers of Structural and Civil Engineering 14, 185–198. doi:10.1007/s11709-019-0591-x.

Melville, B., Yee-Meng, C., 1999. Time Scale for Local Scour at Bridge Piers. Journal of Hydraulic Engineering 125, 59–65. doi:10.1061/(ASCE)0733-9429(1999)125:1(59).

Melville, B.W., 1997. Pier and abutment scour: Integrated approach. Journal of Hydraulic Engineering 123, 125–136.

Melville, B.W., Coleman, S., 2000. Bridge Scour. Water Resources Publications.

Melville, B.W., Sutherland, A.J., 1988. Design Method for Local Scour at Bridge Piers. Journal of Hydraulic Engineering 114, 1210–1226. URL:

https://app.dimensions.ai/details/publication/pub.1057589329, doi:10.1061/(asce)0733-9429(1988)114:10(1210).

Molnar, C., 2019. Interpretable Machine Learning - A Guide for Making Black Box Models Explainable. doi:https://christophm.github.io/interpretable-ml-book/.

Mondoro, A., Frangopol, D.M., Liu, L., 2018. Bridge adaptation and management under climate change uncertainties: A review. Natural Hazards Review 19, 04017023. doi:10.1061/(ASCE)NH.1527-6996.0000270.

Moosavi, V., Talebi, A., Shirmohammadi, B., 2014. Producing a landslide inventory map using pixel-based and object-oriented approaches optimized by taguchi method. Geomorphology 204, 646–656. doi:https://doi.org/10.1016/j.geomorph.2013.09.012.

Naghadehi, M., Samaei, M., M., R., M., N., 2018. State-of-the-art predictive modeling of tbm performance in changing geological conditions through gene expression programming. Measurement 126, 46–57. doi:https://doi.org/10.1016/j.measurement.2018.05.049.

Najafzadeh, M., Barani, G.A., Azamathulla, H.M., 2013. GMDH to predict scour depth around a pier in cohesive soils. Applied Ocean Research 40, 35–41. doi:https://doi.org/10.1016/j.apor.2012.12.004.

Nanni, L., Lumini, A., 2009. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications 36, 3028–3033. doi:https://doi.org/10.1016/j.eswa.2008.01.018.

Naser, M., Kodur, V., 2022. Explainable machine learning using real, synthetic and augmented fire tests to predict fire resistance and spalling of rc columns. Engineering Structures 253, 113824. doi:https://doi.org/10.1016/j.engstruct.2021.113824.

Naser, M.Z., 2021. An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. Automation in Construction 129, 103821. doi:10.1016/j.autcon.2021.103821.

Nasr, A., Kjellström, E., Björnsson, I., Honfi, D., Ivanov, O.L., Johansson, J., 2019. Bridges in a changing climate: a study of the potential impacts of climate change on bridges and their possible adaptations. Structure and Infrastructure Engineering 16, 738–749. doi:10.1080/15732479.2019.1670215.

Naumets, S., Lu, M., 2021. Investigation into explainable regression trees for construction engineering applications. Journal of Construction Engineering and Management 147, 04021084. doi:10.1061/(ASCE)CO.1943-7862.0002083.

Ninić, J., Freitag, S., Meschke, G., 2017. A hybrid finite element and surrogate modelling approach for simulation and monitoring supported tbm steering. Tunnelling and Underground Space Technology 63, 12–28. doi:https://doi.org/10.1016/j.tust.2016.12.004.

Nishimura, A., Tanamura, S., 1989. Study on Integrity Assessment of Railway Bridge Foundation. RTRI Report 3, 41–49. URL: https://cir.nii.ac.jp/crid/1573950401139748608.bib?lang=en.

Olden, J.D., Jackson, D.A., 2002. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154, 135–150. doi:10.1016/S0304-3800(02)00064-9.

Oltean, M., Dumitrescu, D., 2002. Multi Expression Programming.

Ozaeta García-Catalán, R., Martín-Caro, J.A., 2020. Catalogue of Damages in masonry arch bridges. January, UIC. URL: https://shop.uic.org/en/other-documents/9500-catalogue-of-damages-in-masonry-arch-bridges.html.

Pal, M., 2006. Support vector machines-based modelling of seismic liquefaction potential. International Journal for Numerical and Analytical Methods in Geomechanics 30, 983–996. doi:https://doi.org/10.1002/nag.509.

Pala, M., Caglar, N., Elmas, M., Cevik, A., Saribiyik, M., 2008. Dynamic soil–structure interaction analysis of buildings by neural networks. Construction and Building Materials 22, 330–342. doi:https://doi.org/10.1016/j.conbuildmat.2006.08.015.

Park, H., Lee, S., 2011. Evaluation of the compression index of soils using an artificial neural network. Computers and Geotechnics 38, 472–481. doi:10.1016/j.compgeo.2011.02.011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay, 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

Pham, B.T., Tien Bui, D., Prakash, I., 2017. Landslide Susceptibility Assessment Using Bagging Ensemble Based Alternating Decision Trees, Logistic Regression and J48 Decision Trees Methods: A Comparative Study. Geotechnical and Geological Engineering 35, 2597–2611. doi:10.1007/s10706-017-0264-2.

Pizarro, A., Manfreda, S., Tubaldi, E., 2020. The science behind scour at bridge foundations: A review. Water (Switzerland) 12. doi:10.3390/w12020374.

Pizarro, A., Samela, C., Fiorentino, M., Link, O., Manfreda, S., 2017. BRISENT: An entropy-based model for bridge-pier scour estimation under complex hydraulic scenarios. Water (Switzerland) 9, 889. doi:10.3390/w9110889.

Pradhan, B., Lee, S., 2010. Regional landslide susceptibility analysis using back-

propagation neural network model at Cameron Highland, Malaysia. Landslides 7, 13–30. doi:10.1007/S10346-009-0183-2/TABLES/4.

Prendergast, L., Gavin, K., 2014. A review of bridge scour monitoring techniques. Journal of Rock Mechanics and Geotechnical Engineering 6, 138–149. doi:https://doi.org/10.1016/j.jrmge.2014.01.007.

Ramette, M., 1981. Guide d'hydraulique fluviale – Rapport HE/40/81/04 du Laboratoire National d'Hydraulique (in French). Technical Report.

Ran, X., Xue, L., Zhang, Y., Liu, Z., Sang, X., He, J., 2019. Rock classification from field image patches analyzed using a deep convolutional neural network. Mathematics 7. doi:10.3390/math7080755.

Richardson, J.E., Panchang, V.G., 1998. Three-dimensional simulation of scour-inducing flow at bridge piers. Journal of Hydraulic Engineering 124, 530–540. doi:10.1061/(ASCE)0733-9429(1998)124:5(530).

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing 67, 93–104. doi:10.1016/j.isprsjprs.2011.11.002.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206–215. doi:10.1038/s42256-019-0048-x, arXiv:1811.10154.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning Internal Representations by Error Propagation. MIT Press, Cambridge, MA, USA. pp. 318–362.

Saeedi Azizkandi, A., Kashkooli, A., Baziar, M.H., 2014. Prediction of Uplift Pile Displacement Based on Cone Penetration Tests (CPT). Geotechnical and Geological Engineering 32, 1043–1052. doi:10.1007/s10706-014-9779-y.

Samizo, M., 2014. A study on the evaluation of stability of railway bridge piers during the swelling of a river (in Japanese). Ph.D. thesis. Kokushikan University. URL: http://id.nii.ac.jp/1410/00010349/.

Samui, P., 2008. Support vector machine applied to settlement of shallow foundations on cohesionless soils. Computers and Geotechnics 35, 419–427. doi:10.1016/j.compgeo.2007.06.014.

Samui, P., Sitharam, T.G., Kurup, P.U., 2008. OCR Prediction Using Support Vector Machine Based on Piezocone Data. Journal of Geotechnical and Geoenvironmental Engineering 134, 894–898. doi:10.1061/(asce)1090-0241(2008)134:6(894).

Sasidharan, M., Parlikad, A.K., Schooling, J., 2021. Risk-informed asset management to tackle scouring on bridges across transport networks. Structure and Infrastructure Engineering 0, 1–17. doi:10.1080/15732479.2021.1899249.

Shahin, M., 2013. Artificial Intelligence in Geotechnical Engineering: Applications, Modeling Aspects, and Future Directions. Metaheuristics in Water, Geotechnical and Transport Engineering 14, 169–204. doi:10.1016/B978-0-12-398296-4.00008-8.

Shahin, M.A., Maier, H.R., Jaksa, M.B., 2005. Investigation into the robustness of artificial neural networks for a case study in civil engineering. International Congress on Modeling and Simulation, MODSIM 2005 , 79–83.

Shapley, L.S., 1952. A Value for N-Person Games. RAND Corporation, Santa Monica, CA. doi:10.7249/P0295.

Shibayama, T., 2017. Japan's transport planning at national level, natural disasters, and their interplays. European Transport Research Review 9. doi:10.1007/s12544-017-0255-7.

Shirole, A., Holt, R., 1991. Planning for a comprehensive bridge safety assurance program. Transport Research Record , 39–50.

Singh, H., 2021. Beginner's Guide to Low Variance Filter and its Implementation. URL: https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-low-variance-filter-and-its-implementation/.

SNCF, 2005. Cotation des ouvrages d'art Livret A - fondation en site aquatique. Technical Report.

SNCF, 2019. Banque OA-OT (in French).

SNCF, 2020. Patrimoine Informatisé du Génie Civil (PIGC).

Sornel, L., 1872. Bottom of the Sea, New York City, NY: Scribner, Armstrong & Co.

Srivastava, N., Hinton, G., Krizhevsky, A., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Technical Report. doi:10.5555/2627435.

Sumer, B.M., Hatipoglu, F., Fredsøe, J., 2007. Wave scour around a pile in sand, medium dense, and dense silt. Journal of Waterway, Port, Coastal, and Ocean Engineering 133, 14–27. doi:10.1061/(ASCE)0733-950X(2007)133:1(14).

Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A., 2015. Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics 11, 812–820. doi:10.1109/TII.2014.2349359.

Takayanagi, T., Durand, E., Davi, D., Chevalier, C., Cheetham, M., Naito, N., Sanagawa, T., Watanabe, K., 2019. Scour risk management at bridges - A comparison of Japanese and French scoring methodologies URL: https://hal.archives-ouvertes.fr/hal-02359279.

Takayanagi, T., Naito, N., Manom, R., Nunokawa, O., 2018. Evaluation method using score table for identifying bridge piers vulnerable to scouring in Japan, in: Scour and Erosion IX. 1st editio ed.. CRC Press, p. 151.

The pandas development team, T., 2020. pandas-dev/pandas: Pandas. URL: https://doi.org/10.5281/zenodo.3509134, doi:10.5281/zenodo.3509134.

Tocchetti, A., Brambilla, M., 2022. The role of human knowledge in explainable ai. Data 7. doi:10.3390/data7070093.

Todinov, M.T., 2006. Risk-based reliability analysis and generic principles for risk reduction. Elsevier.

Tola, S., Tinoco, J., Matos, J.C., Obrien, E., 2023. Scour detection with monitoring methods and machine learning algorithmsmdash;a critical review. Applied Sciences 13. URL: https://www.mdpi.com/2076-3417/13/3/1661, doi:10.3390/app13031661.

Toth, E., Brandimarte, L., 2011. Prediction of local scour depth at bridge piers under clear-water and live-bed conditions: comparison of literature formulae and artificial neural networks. Journal of Hydroinformatics 13, 812–824. doi:10.2166/hydro.2011.065.

Tubaldi, E., Macorini, L., Izzuddin, B.A., Manes, C., Laio, F., 2017. A framework for probabilistic assessment of clear-water scour around bridge piers. Structural Safety 69, 11–22. doi:https://doi.org/10.1016/j.strusafe.2017.07.001.

UIC, 2022. Railway Statistics Synopsis. Technical Report. International union of railways (UIC). URL: https://uic.org/IMG/pdf/uic-railway-statistics-synopsis-2022.pdf.

Valette, L., Cunillera, A., 2010. Cahiers techniques SYRAH-CE , 93.

Van Leeuwen, Z., Lamb, R., 2014. Flood and Scour Related Failure Incidents at Railway Assets Between 1846 and 2013. Technical Report. URL: https://www.jbatrust.org/how-we-help/publications-resources/risk-analysis/flood-and-scour-related-failure-incidents-at-railway-assets/.

Wakjira, T.G., Alam, M.S., Ebead, U., 2021. Plastic hinge length of rectangular RC columns using ensemble machine learning model. Engineering Structures 244, 112808. URL: https://www.sciencedirect.com/science/article/pii/S0141029621009585, doi:https://doi.org/10.1016/j.engstruct.2021.112808.

Wang, C., Yu, X., Liang, F., 2017. A review of bridge scour: mechanism, estimation, monitoring and countermeasures. Natural Hazards 87, 1881–1906. doi:10.1007/s11069-017-2842-2.

Wang, H.L., Yin, Z.Y., 2020. High performance prediction of soil compaction parameters using multi expression programming. Engineering Geology 276, 105758. doi:https://doi.org/10.1016/j.enggeo.2020.105758.

Wang, H.L., Yin, Z.Y., Zhang, P., Jin, Y.F., 2020a. Straightforward prediction for air-entry value of compacted soils using machine learning algorithms. Engineering Geology 279, 105911. doi:https://doi.org/10.1016/j.enggeo.2020.105911.

Wang, L., Wu, C., Tang, L., Zhang, W., Lacasse, S., Liu, H., Gao, L., 2020b. Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method. Acta Geotechnica 15, 3135–3150. doi:10.1007/s11440-020-00962-4.

Wang, Z., Liu, T., Long, Z., Wang, J., Zhang, J., 2023. Predicting the drift capacity of precast concrete columns using explainable machine learning approach. Engineering Structures 282, 115771. doi:https://doi.org/10.1016/j.engstruct.2023.115771.

Watts, G., Battarbee, R.W., Bloomfield, J.P., Crossman, J., Daccache, A., Durance, I., Elliott, J.A., Garner, G., Hannaford, J., Hannah, D.M., Hess, T., Jackson, C.R., Kay, A.L., Kernan, M., Knox, J., Mackay, J., Monteith, D.T., Ormerod, S.J., Rance, J., Stuart, M.E., Wade, A.J., Wade, S.D., Weatherhead, K., Whitehead, P.G., Wilby, R.L., 2015. Climate change and water in the uk – past changes and

future prospects. Progress in Physical Geography: Earth and Environment 39, 6–28. doi:10.1177/0309133314542957.

Wright, L., Chinowsky, P., Strzepek, K., Jones, R., Streeter, R., Smith, J.B., Mayotte, J.M., Powell, A., Jantarasami, L., Perkins, W., 2012. Estimated effects of climate change on flood vulnerability of U.S. bridges. Mitigation and Adaptation Strategies for Global Change 17, 939–955. doi:10.1007/s11027-011-9354-2.

Xia, Y., Liu, C., Li, Y.Y., Liu, N., 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications 78, 225–241. doi:10.1016/j.eswa.2017.02.017.

Yanmaz, A.M., Altinbilek, H.D., 1991. Study of Time-Depenbent Local Scour around Bridge Piers. Journal of Hydraulic Engineering 117, 1247–1268. doi:10.1061/(asce)0733-9429(1991)117:10(1247).

Younsi, N., 2019. Analyse de Risque des Ponts en Sites Affouillables – Expérimentation et Comparaison des méthodes française et japonaise sur un parc d'ouvrages (Master thesis in French) .

Yousefpour, N., Downie, S., Walker, S., Perkins, N., Dikanski, H., 2021. Machine learning solutions for bridge scour forecast based on monitoring data. Transportation Research Record 2675, 745–763. doi:10.1177/03611981211012693.

Zhang, P., Yin, Z.Y., Jin, Y.F., Chan, T.H., 2020. A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. Engineering Geology 265, 105328. doi:https://doi.org/10.1016/j.enggeo.2019.105328.

Zhang, P., Yin, Z.Y., Jin, Y.F., Chan, T.H., Gao, F.P., 2021a. Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms. Geoscience Frontiers 12, 441–452. doi:10.1016/j.gsf.2020.02.014.

Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D., Zhang, Y., 2022. Application of

machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. Gondwana Research 109, 1–17. doi:https://doi.org/10.1016/j.gr.2022.03.015.

Zhang, W., Wu, C., Zhong, H., Li, Y., Wang, L., 2021b. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geoscience Frontiers 12, 469–477. doi:10.1016/j.gsf.2020.03.007.

Zhao, S., Zhang, D.M., Huang, H.W., 2020. Deep learning–based image instance segmentation for moisture marks of shield tunnel lining. Tunnelling and Underground Space Technology 95, 103156. doi:https://doi.org/10.1016/j.tust.2019.103156.

Zhou, J., Shi, X., Du, K., Qiu, X., Li, X., Mitri, H.S., 2017. Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel. International Journal of Geomechanics 17, 04016129. doi:10.1061/(ASCE)GM.1943-5622.0000817.

Zhu, Z., Liu, Z., 2012. CFD prediction of local scour hole around bridge piers. Journal of Central South University 19, 273–281. doi:10.1007/s11771-012-1001-x.

Zounemat-Kermani, M., Beheshti, A.A., Ataie-Ashtiani, B., Sabbagh-Yazdi, S.R., 2009. Estimation of current-induced scour depth around pile groups using neural network and adaptive neuro-fuzzy inference system. Applied Soft Computing Journal 9, 746–755. doi:10.1016/j.asoc.2008.09.006.

Glossary

Abutment (Culée)	A substructure at the ends of a bridge span supporting its superstructure.
Debris (Embâcle)	Floating or submerged material, such as logs, vegetation, or trash, transported by a stream.
Low flow channel (Lit mineur)	Water flows during the lowest flow conditions. In some cases, the low flow channel may be coincident with the main channel
Main channel (Lit moyen)	The river cross section that carries water during normal (e.g., ordinary) flow conditions.
Masonry (Maçonnerie)	The art and craft of building and fabricating in stone, clay, brick, or concrete block.
Pier (Pile)	The vertical support structures of bridges. Bridge piers are the intermediate supports, whose function is to transmit the forces they receive from the load-bearing elements to the foundations.

Retaining wall (Mur de soutènement)	Structure designed and constructed to withstand lateral pressure of soil or hold back soil materials.
Riverbank (Berge)	The land at either edge of a river. It borders the watercourse and confines the water in the natural channel when the water level, or flow, is normal.
Scour (Affouillement)	Erosion of streambed or bank material due to flowing water.
Sheet piling (Palplanche)	Sections of sheet materials with interlocking edges that are driven into the ground to provide earth retention and excavation support. Sheet piles are most commonly made of steel, but can also be formed of timber or reinforced concrete.
Slope of riverbed (Pente du lit)	The inclination of the channel bottom.
Watercourse (Cours d'eau)	A natural or artificial channel through which water flows.
Wing wall (Mur en aile)	Wing walls are adjacent to the abutments and act as retaining walls.

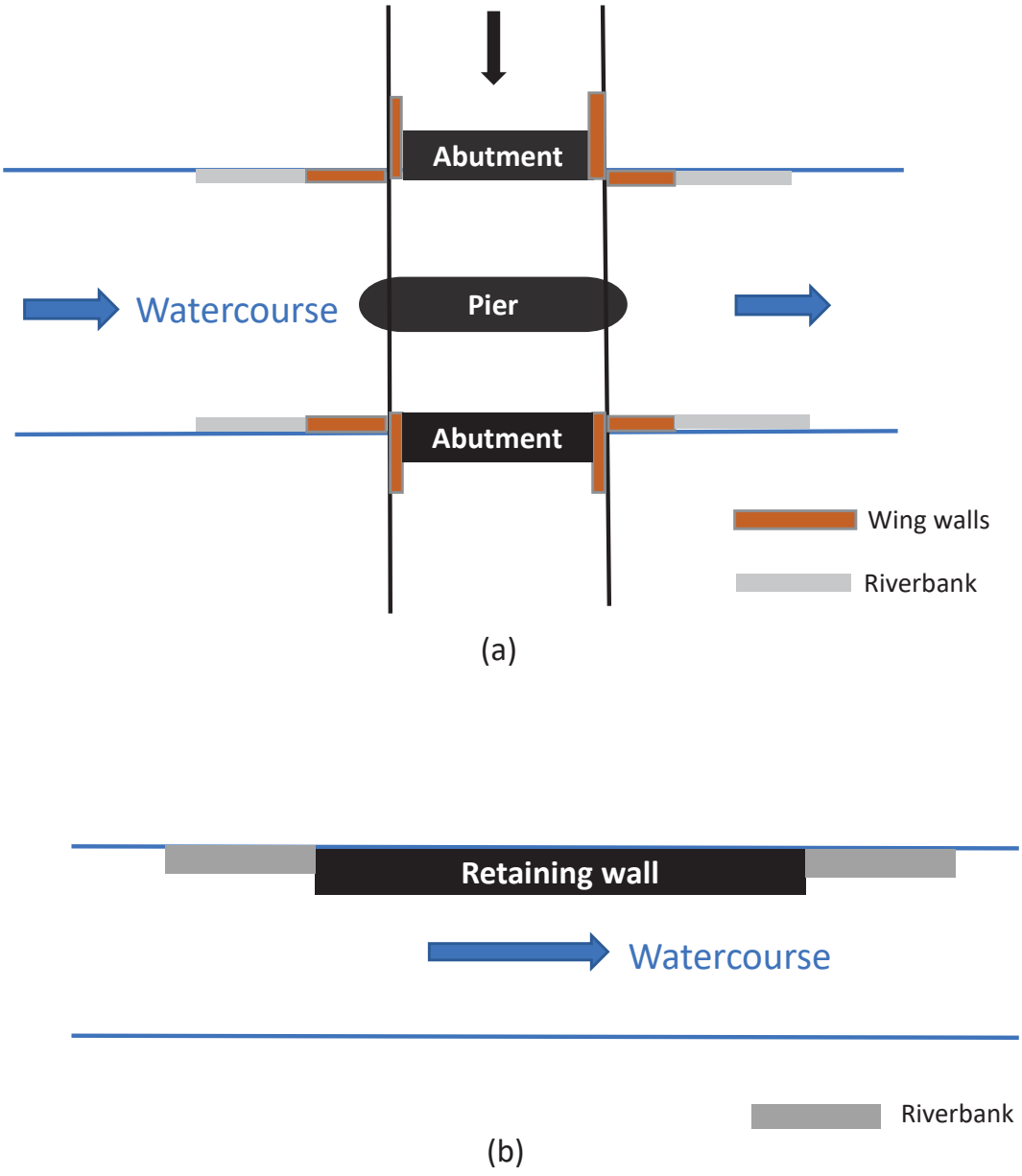
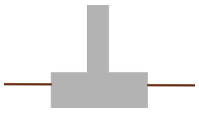
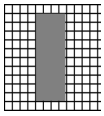
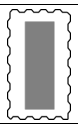
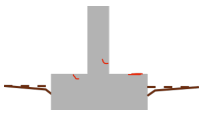
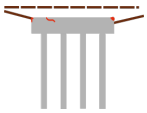
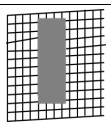
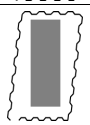
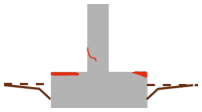
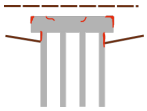
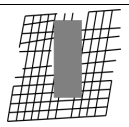
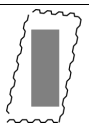
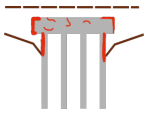
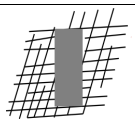
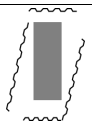


Figure 7.1: Schematic presentations of a bridge crossing watercourse (a) and a retaining wall adjacent to watercourse (b) (top view)

Appendix A

Grade proposition guidance

The purpose of the following table is to help inspectors and engineers make a quick decision and hold the same standard while evaluating in the field. The degraded scenarios in each damage level are generalized. Therefore, it's possible that certain types of damages are not covered here (e.g., settlement, disjoint, dislocation).

Damage level &proposed grade	Foundation		Protection		Comments
	Shallow foundation	Deep foundation	Gabions	Sheet piling	
Minor (9-10)					No obvious damage
Moderate (6-8)					Slight damages at the surface. No influence for the stability of the structure
Extensive (3-5)					Several damages on the surface, the restoration work need to be undertaken
Severe (0-2)					Totally ruined, the circulation could possibly be stopped immediately

Note: If the damages exist on foundation and its protection at the same time, the damage level will be given in accordance with the worst case. For example, the protection has a severe damage level while the foundation has a minor, the final result should be severe.

Figure A.1: Guidance for damage level and proposed grade after inspection

Appendix B

XGBoost model introduction

Extreme gradient boosting (XGBoost) is an ensemble learning algorithm based on tree models.

Suppose a dataset $D = \{(x_i, y_i)\}\,(x_i \in \mathbb{R}, y_i \in \mathbb{R})$ with m features and n examples, x and y represent the input and output data respectively. The predicted output value $\widehat{y_i}$ is calculated as the sum of K additive functions shown in equation (B.1) :

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in \Gamma \tag{B.1}$$

where $\Gamma = \left\{f\left(x\right) = w_{q(x)}\right\}\left(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T\right)$ is the space of decision trees (also known as CART, Breiman et al., 1984) and K denotes the number of decision trees. For a single tree $f_k(x)$, q represents the tree structure that maps the input data to the corresponding leaf node. T is the number of leaves in the tree and w is the leaf weight.

The objective function of XGBoost formulated by Chen and Chen and Guestrin

(2016) is shown in equation (B.2).

$$\begin{aligned} \phi &= \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \\ where\ \Omega(f) &= \gamma T + \frac{1}{2}\lambda \|w\|^2 \end{aligned} \tag{B.2}$$

Here $\sum_{i=1}^{n} l\left(y_i, \widehat{y}_i\right)$ is the loss function which is used to calculate the difference between the actual value y_i and predicted value y_i. Compared with traditional boosting algorithms, XGBoost model adds a regularized term Ω to penalize the complexity of the model in order to avoid overfitting. γ is the cost to have additional leaves. λ is a regularization hyper-parameter and $\|w\|^2$ is the L2 norm of leaf weights.

It's quite difficult to optimize equation (B.2) with traditional approaches in Euclidean space. Therefore, the XGBoost model is trained in an additive approach by using boosting. Boosting is a learning algorithm with which weak learners could be converted to strong ones. By adopting boosting at the j-th iteration, $\widehat{y}_i^{(j-1)} + f_j(x_i)$ represents $\widehat{y}_i^{(j)}$ and equation (B.2) can be transformed to:

$$\phi = \sum_{i=1}^{n} l(y_i, \widehat{y}_i^{(j-1)} + f_j(x_i)) + \Omega(f_j) \tag{B.3}$$

In this case, by adding greedily f_j may help to minimize the objective function and improve the prediction results. To find the optimal value, a second-order Taylor approximation is used to equation (B.4):

$$\phi_j \simeq \sum_{i=1}^{n} \left[l\left(y_i, \widehat{y}^{(j-1)} + g_i f_j\left(x_i\right) + \frac{1}{2} h_i f_j^2(x_i) \right) \right] + \Omega\left(f_j\right) \tag{B.4}$$

where $g_i = \partial_{\widehat{y}^{(j-1)}} l(y_i, \widehat{y}^{(j-1)})$ and $h_i = \partial^2_{\widehat{y}^{(j-1)}} l(y_i, \widehat{y}^{(j-1)})$ are first and second-order gradient statistics of the loss function. After removing the constant term, equation (B.4) can be simplified at step j as:

$$\widetilde{\phi}_{(j)} = \sum_{i=1}^{n} \left[g_i f_j(x_i) + \frac{1}{2} h_i f_j^2(x_i) \right] + \Omega(f_j) \tag{B.5}$$

Define the instance set in leaf t as $I_t = \{i \mid q(x_i) = t\}$, equation (B.5) can be rewritten as follows:

$$\begin{aligned}\widetilde{\phi}_{(j)} &= \sum_{i=1}^{n}\left[g_i f_j(x_i) + \frac{1}{2}h_i f_j^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{t=1}^{T} w_t^2 \\ &= \sum_{t=1}^{T}\left[\left(\sum_{i\in I_t} g_i\right) w_t + \frac{1}{2}\left(\sum_{i\in I_t} h_i + \lambda\right) w_t^2\right] + \gamma T\end{aligned} \tag{B.6}$$

For a given tree structure, define $G_t = \sum_{i\in I_t} g_i$, $H_t = \sum_{i\in I_t} h_i$. Therefore the optimal weight in leaf t is

$$w_t^* = -\frac{G_t}{H_t + \lambda} \tag{B.7}$$

and the corresponding optimal value is

$$\phi_{(j)}^* = -\frac{1}{2}\sum_{t=1}^{T}\frac{G_t^2}{H_t + \lambda} + \gamma T \tag{B.8}$$

Equation (B.8) is a scoring term to measure a tree's performance. A smaller value means the tree is purer and fits the data better.

In reality, it's quite difficult to enumerate all the possible tree structures q because the tree is grown greedily. As a result, the first only has a single leaf and it will iteratively add branches. After each split, the gain in the loss reduction is calculated as follows:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{B.9}$$

Equation (B.9) can be obviously divided into four parts. $\frac{G_L^2}{H_L+\lambda}$, $\frac{G_R^2}{H_R+\lambda}$ and $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ are score for the left child, score for the right child and score before splitting respectively. γ is the regularized hyper-parameter by introducing additional leaf. If $Gain < 0$ the splitting will be stopped. More detailed information about XGBoost can refer to the research work of Chen and Guestrin (2016).