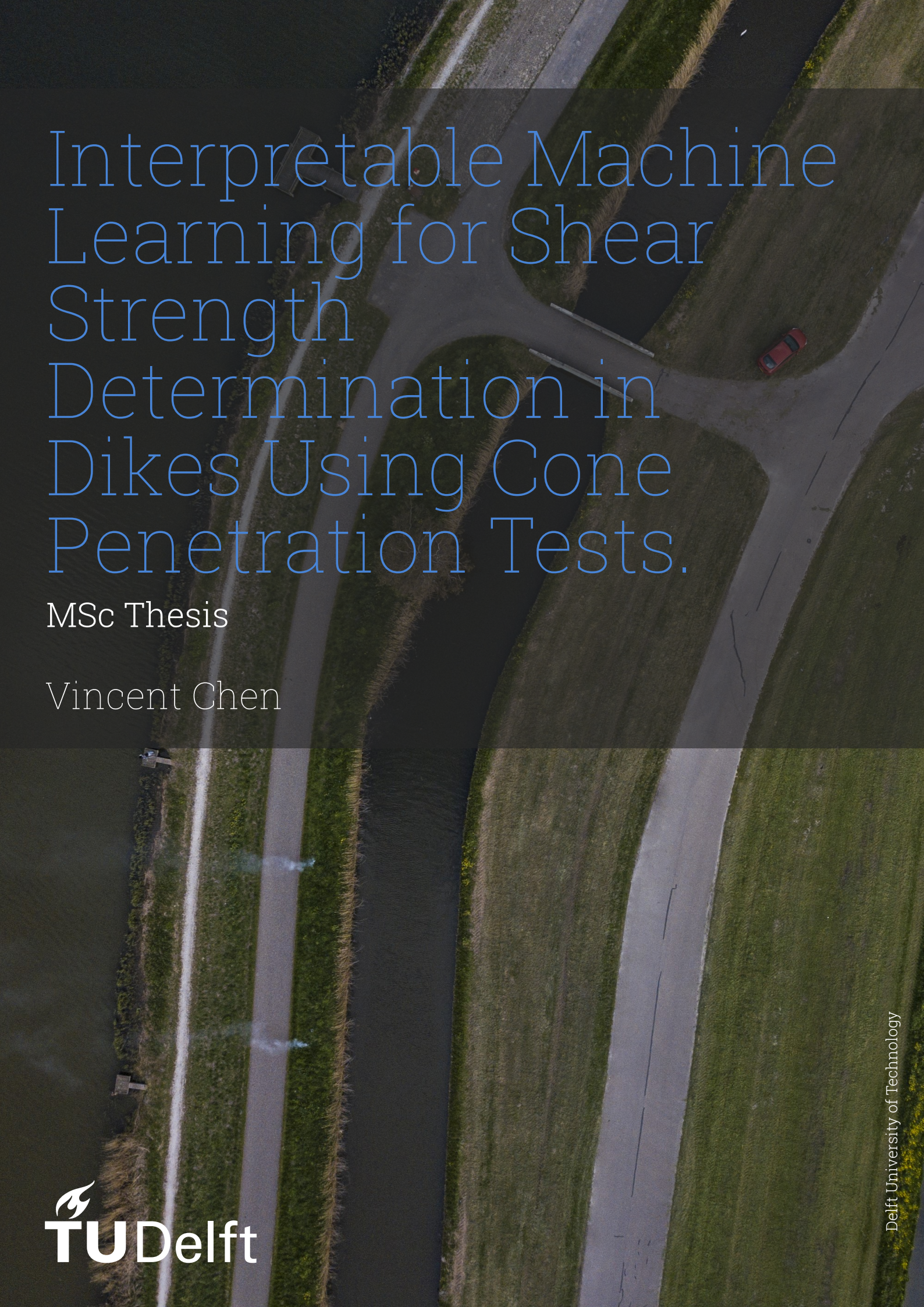


Interpretable Machine Learning for Shear Strength Determination in Dikes Using Cone Penetration Tests.

MSc Thesis

Vincent Chen

Delft University of Technology

TU Delft

Interpretable Machine Learning for Shear Strength Determination in Dikes Using Cone Penetration Tests.

MSc Thesis

by

Vincent Chen

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on March 26th, 2025 at 11:00.

Student number: 5169003

Date: March 19th, 2025

Thesis committee:

Prof. dr. ir. C. Jommi,	TU Delft, chair
Dr. ir. S. Muraro,	TU Delft
Dr. R. Taormina,	TU Delft

TU Delft

Preface

This thesis has been a great opportunity to combine strong geotechnical knowledge with flashy new artificial intelligence methodologies, which is something I have grown a passion for over my five and a half years of studying at TU Delft. Through various projects, I have now layed a foundation to build further on during my career. Thank you to my supervisors Prof. Dr. Cristina Jommi, Dr. Stefano Muraro, and Dr. Riccardo Taormina for providing me this opportunity, taking their time to guide me throughout the project, and supporting me during the ups and the downs in the past months. Their expertise, enthusiasm, and professionalism continuously inspires me to grow as an engineer and person.

I am thankful for my family, from whom I have always felt support and encouragement, helping me overcome the challenges during the past months. Thank you to the amazing people I have met during the years of studying for the times we have spent together in lecture halls, study rooms, and abroad. And finally, of course, thank you to my wonderful friends, for being there for me and keeping me grounded, even when my mind was wandering off thinking about the best strategy for my next models.

Vincent Chen
Delft, March 2025

Summary

In the Netherlands, the proper assessment of a dike's slope stability is an area of increased interest, as large stretches of dikes need to be reinforced according to current assessment methodology. To minimize costs, both in terms of money and labor, it would be preferable to reduce the amount of dike reinforcement work.

In current assessment methodology, field investigation is commonly performed in summer, when the dike is in a drier state than it is in the winter, but calculations are done for the governing situation, which assumes wet, winter conditions. As a result, field measurements should be translated, for which no grounded guideline exists yet. Thus, to properly assess a dike's seasonally dependent strength, a new shear strength determination methodology should be developed.

Many research has been performed regarding the cone resistance measurement, q_c, of a cone penetration test (CPT). Often overlooked, however, is sleeve friction, f_s, which could provide valuable information. Therefore, this thesis investigates the potential use of this parameter, in combination with available weather data, to determine shear strength. The following research question is answered:

How can machine learning models be used in the determination of shear strength of unsaturated layers in a dike using cone penetration test and weather data?

The main approach to answering this question involved neural networks and random forests. Through SHAP feature analysis, the working of the machine learning models was interpreted, which helped improve models iteratively through feature engineering. The best performing models used engineered CPT and weather features to predict undrained shear strength directly, and achieved R^2 scores over 0.7, although the models were not yet able to transfer knowledge from one dike to another. Prediction of volumetric water content proved to be difficult for the models, since no accurate test performance was achieved after the removal of depth as a feature, which was done in order to reduce the risk of overfitting.

As a result of iterative improvement and interpretation of the machine learning models, a better understanding of CPTs for the prediction of volumetric water content and undrained shear strength was achieved. It was found that sleeve friction was the most important feature for neural networks, both in the prediction of water content and undrained shear strength. These neural network models made little use of cone resistance, but were able to improve when friction ratio (sleeve friction divided by cone resistance) was added in the input layer. Random forests were found to work differently, as they were able to combine CPT and weather features, and have them interact with each other to make similar predictions to the neural networks.

Through further analysis of feature interpretation, it was found that a strong, positive, linear correlation was found by the neural networks between sleeve friction and the predicted shear strength. The friction ratio was found to introduce non-linearity to the shear strength predictions. Weather features were used to a lesser extent, indicating that the CPT parameters capture seasonal effects to a sufficient extent. In conclusion, sleeve friction measurements were found to be an important parameter for shear strength determination, with improved machine learning predictions when compared to conventional methodologies.

Contents

List of Figures

List of Tables

1

Introduction

1.1. Background

Currently in the Netherlands, slope stability analyses of dikes are performed assuming fully saturated conditions for both the subsoil and the dike body (van Duinen, 2021), which represents the state at which the dike is most likely to suffer a slope stability failure. Shear strength is an important factor in this assessment, which is commonly determined through field (e.g., CPT) and laboratory tests. However, since field tests are typically performed during summer, when a significant part of the dike may be unsaturated, a discrepancy arises between the saturation conditions during testing and the wet state assumed for assessment. A reduction in the shear strength estimated from summer tests is needed to account for wetter conditions, characterized by higher water content and lower suction. The method of reducing the strength estimated from summer measurements is not yet well-defined, potentially leading to either an overconservative or unsafe assessment or design.

The part of the dike that is unsaturated in the summer and becomes more or fully saturated during winter (i.e., the part with a seasonally dependent water content) is referred to as the initially unsaturated layer. Figure 1.1 shows a schematic drawing of this layer. In the east of the Netherlands, this layer can even extend into the subsoil during periods of drought. Assessment of shear strength in unsaturated soils is more challenging than the saturated counterpart due to the presence of both air and water in the pore space. To account for this in an effective stress approach, an estimate is required of the degree of saturation and suction to describe the state of the soil, in addition to the void ratio for saturated soil. For saturated conditions, the typical seasonal approach in the Netherlands for dikes is based on a total stress approach, although an effective stress approach would lead to a more appropriate strength determination. The latter would however require knowledge of pore pressure build up in the field, for example by performing a fully hydro-mechanical coupled analysis. Considering the complexity of incorporating such an analysis, a total stress approach has been preferred in practice thus far.

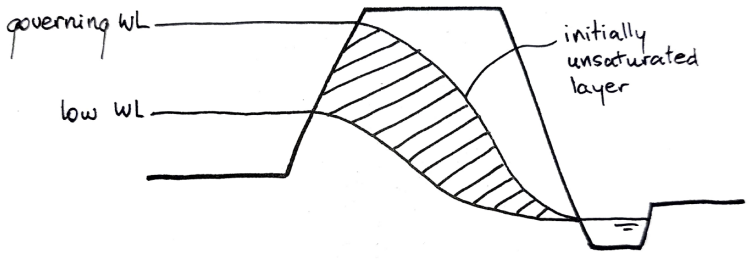


Figure 1.1: Schematic dike cross section, visualizing the initially unsaturated layer defined by the governing and low water levels.

Since low permeable, initially unsaturated soil layers are assumed fully saturated in slope stability analyses, they are commonly assessed using an undrained approach, though some waterboards apply a drained approach. The difference is that the former assumes a constant water content and does consider build up of pore water pressure, while the latter assumes a varying water content and does not consider pore water pressure development. A consequence analysis carried out by Arcadis (2020) shows that the choice of using a drained approach over an undrained approach could lead to a stability safety factor that is 1.5 times higher. In stability assessment, this could lead to a probability of failure up to 10,000 times lower, and in design, it could mean that there is no more need for a stability berm, if this was the case before.

Since even for saturated conditions no unique assessment methodology is agreed upon, and taking into account the additional challenge inherent to unsaturated soil mechanics, the need for an updated guideline arises, which should feature a consistent and clear approach to assessing the slope stability of dikes with an initially unsaturated layer. TU Delft and Deltares are involved in the research needed to achieve this goal, which investigates the seasonal variation of shear strength in dikes. For this, two primary dikes in the east of the Netherlands were selected as test locations, where field tests were conducted and samples were retrieved for laboratory experiments. The initially unsaturated layer at these locations consisted of mostly clayey silts. Cone penetration tests (CPTs) were conducted on site, which is one of the most common techniques used in site characterization over the past 50 years. The field tests were repeated at close locations over time, making analyses of time dependent processes, such as seasonal variations, possible.

The cone resistance measured by CPTs, q_c, features in a wide range of empirical correlations which are useful in practice, among which the determination of the undrained shear strength. The formulation of the relationship between q_c and the undrained shear strength, s_u, that is currently applied in the guidelines is:

$$s_u = \frac{q_c - \sigma_v}{N_k} \tag{1.1}$$

Where N_k is the empirical cone factor.

Research shows that a dependency of q_c on moisture content and matric suction exists (Miller et al., 2018). A clear relation cannot yet be defined however, as significant scatter is still observed. Thus, more research is needed to be able to better describe this relation, in particular with regard to shear strength of unsaturated layers. One possibility, is to additionally make use of the sleeve friction, f_s, which is measured in parallel with q_c and could reduce uncertainty.

1.2. Objective, Scope and Research Questions

This research project aims to investigate to what extent the sleeve friction can be used in CPT-based estimation of undrained shear strength, s_u. In practice, f_s is rarely used in the evaluation of the undrained shear strength, though it is hypothesized to hold valuable information related to the state of the soil, which is not captured by q_c. In this investigation, it is important to obtain an increased understanding on how to interpret f_s and other relevant influence factors. This investigation will be performed through the use of machine learning models, offering a data-driven approach that is novel to this specific problem.

Focusing on primary dikes in the Netherlands, this research uses previously collected data by Deltares. This data consists of measurements of (among other) CPTs, field vane tests (FVTs), water content sensors, and tensiometers, situated in initially unsaturated layers and performed over a period of 1 year.

The main research question is as follows:

How can machine learning models be used in the determination of shear strength of unsaturated layers in a dike using cone penetration test and weather data?

The following sub-questions will contribute towards answering the main research question:

1. How are CPTs typically interpreted and what relations to seasonal influence have been found?
2. To what extent is seasonal influence present in the dataset, consisting of CPT, FVT and weather data?

3. Can machine learning models be used to make predictions regarding the hydro-mechanical state of the dike?
4. How do machine learning models interpret CPT and weather-related features?
5. To what extent can machine learning models be improved through knowledge-based feature engineering?

1.3. Approach

This research will feature an interpretation and analysis of the data provided by Deltares, focusing on the seasonal dependency of measurements. Based on the data analysis, an assessment regarding the usability of the data for machine learning models is made, and it is chosen what input and output layers the machine learning models should consist of.

After performing the data analysis, machine learning models are built using two different methods: neural networks and random forests. In past research, these machine learning methods have proven especially useful in non-linear problems, and would be expected to deal well with the complexity inherent to the problem. The application of a machine learning model should aim to incorporate as much physical knowledge of the problem. Through SHapley Additive exPlanations (SHAP) feature analysis, the models' interpretations of features are analyzed. Using this additional knowledge of feature interpretation, the features in the input layer are engineered to build new models, aiming to improve performance. Through this iterative process, feature interpretation is extensively analyzed, allowing for new insights on how to use those features. In the end, the best performing models are tested on their ability to generalize in a transferability test.

1.4. Structure of the Thesis

The thesis will start with a literature review, providing theoretical background knowledge for understanding the different aspects of the problem and their complexity. The methodology for the thesis is then explained. Next, the available data will be analyzed. Next, the test sites are described and the data are analyzed. The results are given for the machine learning modeling results.

2

Literature Review

In this Section, the findings in the literature review are presented. The purpose of this section is to provide the reader with necessary knowledge on the topic, as well as to substantiate the motivation for this thesis.

2.1. CPT Interpretation in Saturated Soils

Cone penetration tests are used in the site investigation for a wide variety of engineering projects, as these tests are relatively quick, cheap, and easy to perform, and they offer a near continuous profile. A traditional piezocone penetration test, CPTu, measures cone resistance, q_c, sleeve friction, f_s, and pore water pressure at different locations along the piezocone, u_i. Figure 2.1 displays a schematic representation of a CPTu device, including the locations of the pore water pressure measurement devices. According to ASTM standards, friction sleeves have a height of 134, 145, or 164 mm.

As stated in Lunne, Robertson, and Powell (1997), the CPT site investigation serves three purposes:

1. to determine sub-surface stratigraphy and identify materials present;
2. to estimate geotechnical parameters;
3. to provide results for geotechnical design.

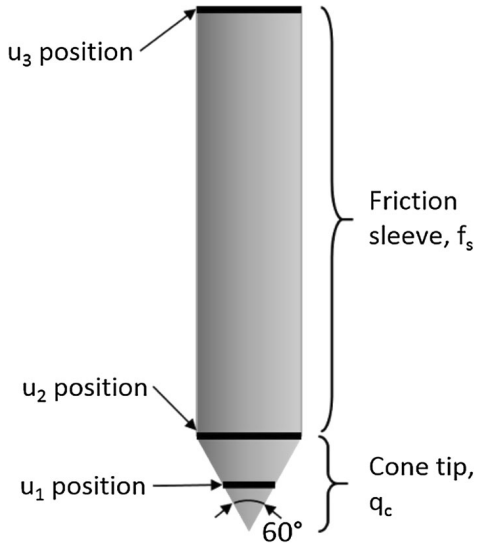


Figure 2.1: Schematic representation of CPTu device, adapted from Ceccato and Simonini (2017).

When dealing with soft clays and silts with high pore pressures and low cone resistance measurements, it is thought to be necessary to account for the unequal end area of the cone and sleeve of the penetrometer. Campanella et al. (1982) thus advised to apply the following correction:

$$q_t = q_c + u_2 * (1 - a) \tag{2.1}$$

Where a represents the net area ratio, often approximated by $a = \frac{d**2}{D**2}$, where d is the diameter of the load cell support and D is the cone diameter. Similarly, to account for an unequal end area between the bottom and top of the friction sleeve, Konrad (1987) advised to also apply a correction to the sleeve friction measurements:

$$f_t = f_s - (1 - \beta b)cu_2 \tag{2.2}$$

Where $b = \frac{A_{st}}{A_{sb}}$, the ratio between the cross sectional area of the top and the area of the bottom of the friction sleeve, $c = \frac{A_{sb}}{A_s}$ is the ratio between the cross sectional area of the base and at the mantle of the friction sleeve, , and $\beta = \frac{u_3}{u_2}$. However, this sleeve friction measurement correction rarely is used in practice, as the difference between f_s and f_t is negligible when the cone has an equal end area sleeve, as argued by Robertson (1990).

In saturated conditions, the distinction between an undrained and a drained analysis is often made. In an undrained analysis, constant volume shearing is assumed, causing development of pore water pressure, and it is applied when dealing with fine soil types, for example clay. In a drained analysis, zero excess pore water pressure is assumed, as water can freely move in the pore space. As fine grained soils are often found in Dutch dikes, undrained analyses are commonly applied. CPTs can then be used to estimate the undrained shear strength through the following relationship:

$$s_u = \frac{q_t - \sigma_{vo}}{N_{kt}} \tag{2.3}$$

Where σ_{vo} is the initial vertical stress and N_{kt} the normalized cone factor. Teh and Houlsby (1991) applied an analytical method to investigate the factors which influence N_{kt} in clay. It was found, that N_{kt} depends on the soil rigidity, in-situ stress conditions, cone roughness and shaft roughness. Interestingly, the horizontal stresses were found to have a greater influence than the vertical stresses. The realization should be made that the rigidity index, I_R, is related to the Over Consolidation Ratio (OCR) and Plasticity Index (PI) (Keaveny and Mitchell, 1986).

As is the case with undrained shear strength determination, in engineering practice, q_c is often used as a strength defining parameter. f_s, on the other hand, sees limited implementation in correlations used in guidelines. The reason for this is that there is a common consensus that sleeve friction measurements are less accurate than cone resistance measurements. Lunne and Andersen (2007) stated the following reasons for this:

- Pore pressure effects on the ends of the sleeve;
- Tolerance in dimensions between the cone and sleeve;
- Surface roughness of the sleeve;
- Load cell design and calibration.

The effect of surface roughness on CPT sleeve friction measurements was investigated by DeJong et al. (2001). On average, friction sleeve measurements resulting from a CPT with a textured friction sleeve were 1.8-1.88 times higher than those resulting from a CPT with a smooth friction sleeve. Figure 2.2 shows the difference in measured f_s profiles for smooth and textured friction sleeves.

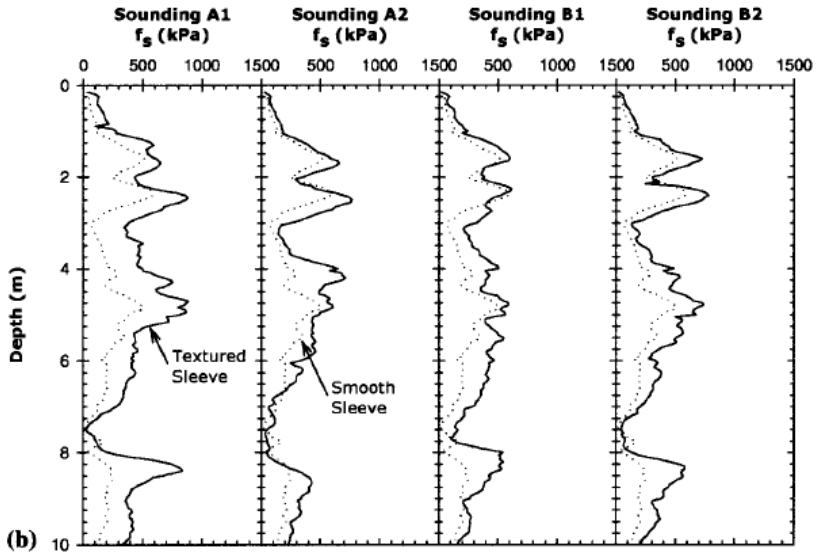


Figure 2.2: f_s Profiles obtained with textured friction sleeves, adapted from DeJong et al. (2001).

Though uncertainties exist in sleeve friction measurements, it should be noted that repeatable sleeve friction measurements are obtainable, as mentioned in for example Robertson (2009). Robertson (2009) also makes use of sleeve friction measurements in predicting soil behaviour directly from CPTs through the soil behaviour type (SBT) chart, which was originally introduced in Robertson (1990). Figures from Farrar et al. (2008) are also presented in this paper, which indicate a relationship between sleeve friction measurements and the remolded shear strength measured by vane shear tests.

2.2. CPT Interpretation in Unsaturated Soils

When interpreting CPTs performed in unsaturated soils, an added level of complexity is added. In saturated conditions, one could make the decision to either perform a drained or an undrained analysis. Such a discrete drainage assumption does not suffice for unsaturated soils, however, since complex coupled processes exist between pore air pressure, pore water pressure and soil volume changes. Affecting suction, these processes are of significant influence on CPT measurements.

In unsaturated soil, u_2 is not obtainable, as the pore pressure sensors need to be fully saturated, and due to the unsaturated condition, the sensor cavitates. Therefore, the definition of the relationship between s_u and q_c changes slightly:

$$s_u = \frac{q_c - \sigma_v}{N_k}$$

Though Equation 2.3 and Equation 2.2 are similar, N_{kt} and N_k are not required to be similar, since measured q_c values are substantially higher when higher suction is present, which pairs with lower moisture content. One paper which demonstrates this is Miller et al. (2018), where CPTs were performed over a period of time, together with moisture content measurements. A seasonal variation of both q_c and f_s was observed, where measurements increased for decreasing moisture content. Interestingly, at one site, the friction ratio, F_R, seemingly increased with increasing moisture content. Measurements from this site are given in 2.3.

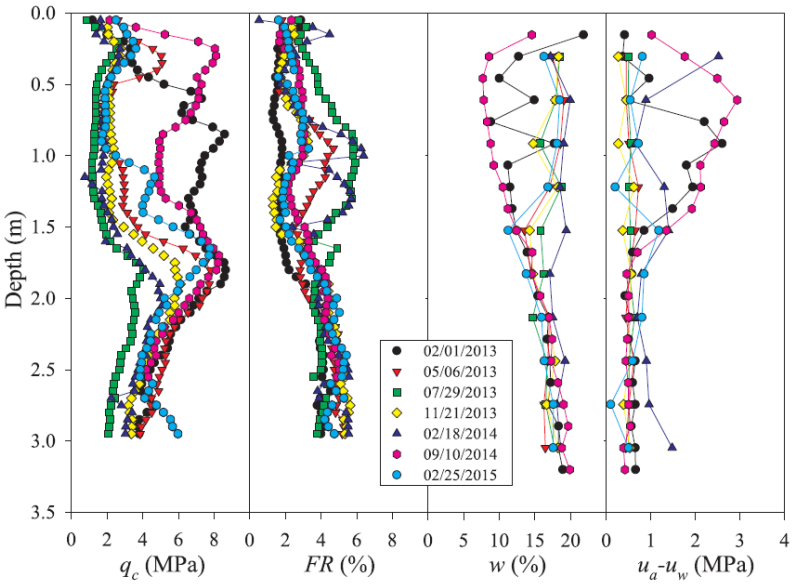


Figure 2.3: CPT tip resistance, friction ratio, water content and matric suction under seasonal variation. Adapted from Miller et al. (2018)

Miller et al. (2018) also found that the dependency of q_c on suction was greater for significantly plastic soils. J.J.M. Powell (1988) stressed the importance of plasticity, and demonstrated a dependency of N_k on the plasticity index, PI. This relationship is not found for all soil types, however. In Zein (2017), for example, no correlation is found between PI and N_k for fine grained Sudanese soils, but a definable relationship was found between OCR and N_k. The paper also stressed the dependency of N_k on moisture condition and degree of stiffness, though no direct relationship was found between $q_c - \sigma' \sigma_v$ and N_k.

No methods regarding the inclusion of suction in CPT interpretation is yet agreed upon. Giacheti et al. (2019) is another paper which investigates the importance of suction on cone resistance measurements. It is demonstrated that by including a suction-based reduction of q_c to translate between dry-season and wet-season measurements, similar normalized Q_{tn} profiles result. It is also stressed that two distinct regions exist on the soil water retention curve, as seen in Figure 2.4. In region 1, the water content greatly varies with small variations of suction. In region 2, the opposite occurs, as large changes in suction exist for small changes in water content.

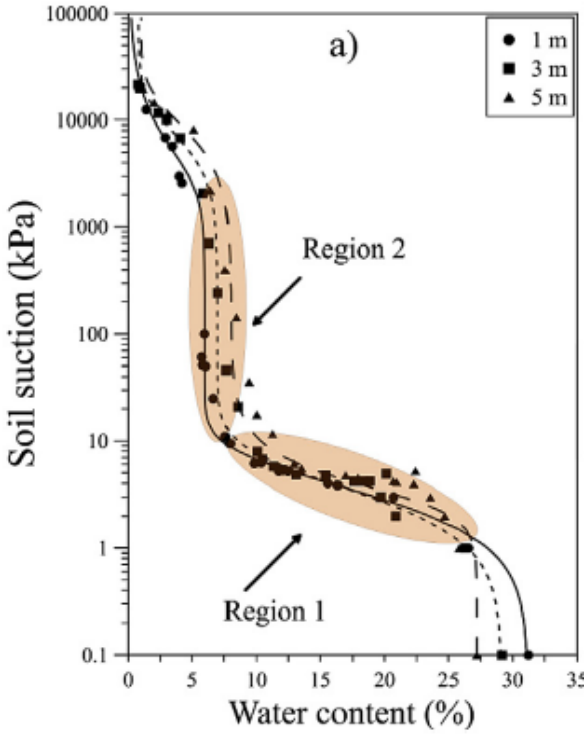


Figure 2.4: Soil water retention curve regions. Adapted from Giacheti et al. (2019).

Multiple publications (e.g., Miller et al. (2018), Giacheti et al. (2019)) have highlighted that the water content mostly varies in the top meters of a soil column, and that there is a depth from which no more water content, and thus suction, variations exist. The top part of a soil column which has a varying water content is referred to as the initially unsaturated layer by van Duinen (2021). Taking into account Figure 2.4, it can be stated that suction values can vary greatly when the soil is in a state where the water content is close to the residual water content.

2.3. Seasonal Variability

The undrained shear strength (s_u) of soils is highly variable due to factors such as natural heterogeneity, stress history, and seasonal influences. Studies have shown that s_u can exhibit significant spatial and depth-dependent variations, particularly in soft clays and organic soils Lunne, Robertson, and Powell (1997) and Phoon and Kulhawy (1999). Therefore, field tests such as CPTs and FVTs are also highly variable, which is partly caused by seasonal variability. This Subsection aims to provide knowledge of how this seasonal variability works and might be noticed in the dataset.

Almasoudi et al. (2023) presents laboratory tests where Kaolin clay was sheared against a metal surface. The adhesion effects were investigated. In Figure 2.5, the interface adhesion results are plotted against moisture content. As moisture content increased, it was observed that adhesion reduced, especially past a moisture content of 30%. The research related this to the occurrence of matric suction

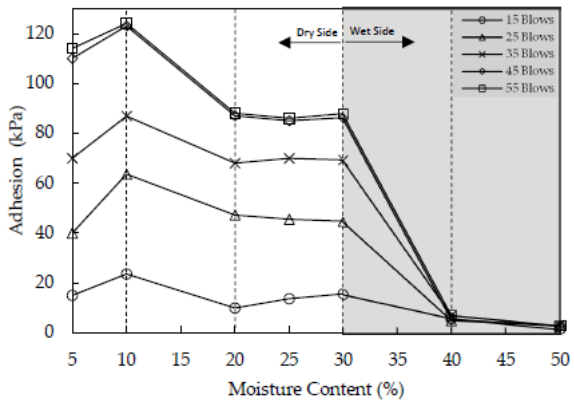


Figure 2.5: Interface adhesion of kaolin at different moisture contents and densities.

Figure 2.6 shows how the adhesion behavior differed for the dry side versus the wet side. One of the main conclusions of the research was, that a linear relationship exists between adhesion and dry density on the dry side, due to a larger surface contact. Almasoudi et al. (2023) recommended further research to consider different types of soil and mixtures, and to perform tests under different temperatures and shear rates.

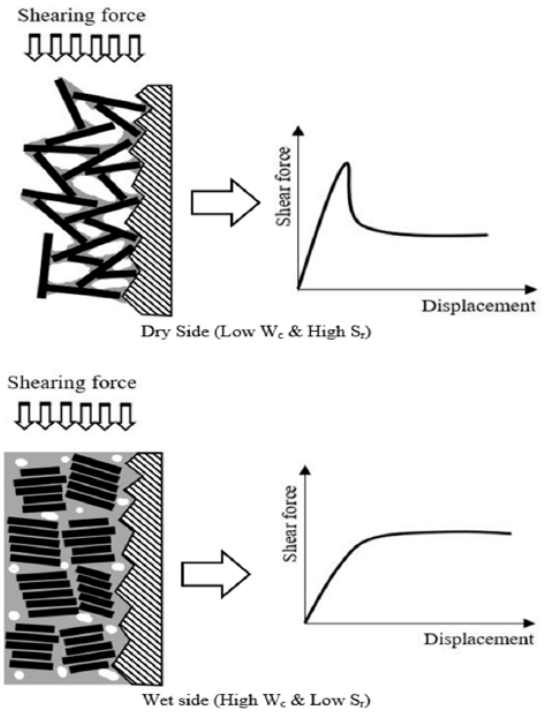


Figure 2.6: Interface adhesion behavior for dry side versus wet side.

2.4. Machine Learning Application in Geotechnical Engineering

In Elsawy et al. (2022), an approach to predict undrained shear strength from a set of 8 soil parameters is described and tested. Highly accurate test performance was recorded, with an R^2 score of 0.96. The input features were as follows:

- Moisture content (θ)
- Specific gravity (G_s)

- Void ratio (e)
- Dry unit weight (γ_d)
- Liquid limit
- Plasticity indexes
- Consistency indexes
- Pocket penetration shear

This research thus indicated a correlation between these features and undrained shear strength.

Shao et al. (2023) discusses different applications of machine learning in geotechnical engineering. It concludes, that support vector machines, neural networks and decision trees are most popular. The paper reviewed some of the applications, and concluded that machine learning methods have been successfully applied in the field of geotechnical engineering for both classification and regression problems.

Zhang et al. (2021) reviewed the application of deep learning methods in geotechnical engineering practice. They discussed the use of various advanced deep learning methods, such as feedforward neural networks (FNN), recurrent neural networks (RNN), convolutional neural networks (CNN), and long short term memory models (LSTM). It concludes, that FNNs are most popular in specific applications. RNNs are most suitable for time series problems, with LSTM models excelling at long term predictions. CNNs are best used for image processing.

2.5. Limitations and Research Gap

Research has investigated the influential factors on CPT measurements. A known empirical relationship between q_c and s_u was found, namely through N_k, though no generic definition of N_k has been agreed upon. N_k is found to depend on multiple factors, such as hydraulic state, plasticity and OCR. From the constitutive framework for unsaturated soils, some of these factors can also be reasoned to be reflected in CPT sleeve friction measurements, f_s. However, it has not yet been investigated whether information captured in f_s can bypass the need for N_k. This is partly due to limited understanding of f_s and its influence factors. No research to what extent f_s can be used in shear strength determination.

Furthermore, the influence of moisture content to interface adhesion was researched. It was found that there a different relationship between moisture content and adhesion exists between a dry and wet state. Similarly, it was found that two regions can be located on the soil water retention curve. The two regions could be related to the two states. The fact that soil behavior is different for different water content values could be one of the limiting factors in translating dry, summer measurements to wet, winter predictions.

Using soil properties, in combination with information about the moisture content, it was previously found that accurate predictions of s_u can be made. However, in practice, many of the used features in Elsawy et al. (2022) are unknown. Hence, the application of CPT data, which are often performed in geotechnical site investigation, would be a more practical approach to predicting geotechnical properties in a dike. Though, during the literature review, little research was found regarding the use of f_s for peak shear strength prediction. Since data-driven machine learning models have been found to be effective in the prediction of geotechnical parameters, these types of models could be used to explore the use f_s, and by extension q_c, depth, and other relevant and available parameters.

3

Methodology

This Chapter explains the methodologies used for the data analysis (Chapter 4) and machine learning modeling (Chapter 5). Chapter 4 gives an overview of the available data and presents an initial data analysis. The results of this analysis are useful in the decision making regarding input and output features for the models that are built in Chapter 5.

3.1. Data Analysis

Chapter 4 starts off with a description of the test sites: the Maasdijk in Oijen and the IJsseldijk in Westervoort. At these two sites, it is important to know to what extent the in-situ data are reflecting seasonal variability. Seasonal influence is thought to be caused by weather effects, thus precipitation and evaporation data of nearby weather stations are reported. Additionally, the water content sensor data can help understand the subsurface effects of seasonal influence. The seasonal effect on the mechanical properties of the soil are then analyzed, by presenting and analyzing measurements throughout the year of CPTs, FVTs, and the tensiometers. In this way, a first observation can be made regarding the storage of seasonally dependent information of each of the measurements, which is useful for selecting meaningful features of the machine learning models.

Next, in order to provide a better understanding on how CPT data can be the most valuable as model input, CPT data is plotted against the two main parameters of interest: water content (θ) and undrained shear strength (s_u). An initial conclusion can be made regarding the use accuracy of models that use CPT data to predict those parameters of interest.

3.2. Machine Learning

Two machine learning methods are used in this thesis: artificial neural networks and random forests. Both models are known to perform well for non-linear, complex problems, such as this thesis'. Neural networks are popular in academic use, though they often need a large dataset in order to find patterns and generalize well. Meanwhile, random forests are expected to perform better for tabular, small datasets, but are therefore also more prone to overfitting, since variations and noise of a small dataset can influence the decision trees, leading to worse generalization. In addition, random forests are considered to be easier to interpret than neural networks, since the decision-making process is more transparent than that of a neural network, which is often considered a black box. This quality of a random forest is considered important.

The reason for using two different methods, is to be able to compare the suitability of the data for different methodologies, and to conclude on a robust model. Additionally, as the methodologies work differently, it is expected that features might be also be used differently. Therefore, a better understanding results regarding the feature importance, and the results of both models can complement each other. Elaboration on the methods can be found in Section 3.2.2 and 3.2.3.

Two different types of models will be created in parallel, which have different target parameters. They are as follows:

- Models type A - Predicts volumetric water content (θ)
- Models type B - Predicts peak shear strength (s_u)

Although the two are thought to be related to each other, the data scarcity has lead to the splitting into two different model setups. Further elaboration can be found in Section 4.6. Each of the two model types will be the setup for a neural network and for a random forest.

3.2.1. Modeling Pipeline

The pipeline used for machine learning modeling and feature importance analysis is as follows:

- **Selecting input features:** the input layer is constructed.
- **Preprocessing data:** features are normalized and split into training/validation and test data.
- **Model training:** models are calibrated using training data. Hyperparameter tuning is performed to ensure the best model performance.
- **Model testing:** model performance is assessed using test data.
- **Feature importance analysis:** SHAP feature importance is analyzed to understand the model's working.

At the end, the working of a model is concluded, and the next step to improve performance is contrived. The pipeline is repeated using new features, at which point new models are built. In the end, the feature analyses of different models allows for better understanding of the data and its physical meaning. Below is a further explanation of the pipeline and some of the decisions made.

Based on the data analysis, a set of input features is chosen. In order to ensure equal initial feature contribution, all features are normalized within its own dataset by MinMax normalization, which means that each value will be scaled between 0 and 1; 0 being the dataset's minimum value and 1 being the dataset's maximum value. This way, each feature has equal chances to affect the model. Additionally, a bias term is always included in the input layer, which is consistently equal to 1. The bias term is added in order to prevent underfitting by introducing an initial offset (Bishop, 2006).

Of the dataset, 15% is kept for the test dataset. The other 85% is used for training and validation, which uses k-fold cross validation. The training set is used to calibrate model parameters, the validation set is used to tune hyperparameters and to stop training to prevent overfitting, and the test set is used to give an assessment of the model's performance on an unseen set. The 85/15 ratio is a conventional ratio in machine learning practices, as it is a balanced way to deal with the purpose of each subset (Bishop, 2006).

Neural network models are then trained using K-fold validation with 5 folds. This entails that the training/validation set is split in an 80/20 ratio for training and validation data. This is done 5 times, such that a different 20% of the test is used for validation each fold. K-fold validation is performed in order to maximize the use of the data, which is especially valuable in data-scarce problems. This results in 5 differently calibrated models. The final prediction is the average of the predictions of each model. Using validation performance, hyperparameters are tuned to minimize validation loss. It was chosen to not perform the same K-fold validation for the random forest models, since the random forest already combines many different decision trees. Performing K-fold validation could, however, lead to improved tuning of the hyperparameters, since it combines models with different bootstrapping, improving generalization of the model.

Models are then tested using the test dataset, which was kept unseen from the model so far, and performance is assessed using the R^2 score and mean squared error (MSE) of the predictions, quantifying both the correlation and error.

Using the training/validation set, the importance of features are analyzed. For both methods, this will be done using SHapley Additive exPlanations (SHAP), which is a machine learning method that helps explain models' output with regard to its input. Additionally, it can be used to understand how input

features work together. The SHAP methodology is explained further in 3.2.4. Essentially, it can be considered to be an extensive sensitivity analysis for machine learning methods.

Through this feature importance analysis, the models' functioning is interpreted. This interpretation can then be used in order to improve models' performance by changing the input layer, possibly resulting in performance improvement through two ways: reduction of the dimension of the input layer, and engineering of the features. By reducing the dimension of the input layer, the curse of dimensionality is avoided, and computational expense is saved. The curse of dimensionality refers to the challenges that arise as the number of features increases. Large input layers cause data points to become sparse, making it harder for models to generalize, leading to overfitting and increased computational costs. The feature importance analysis will allow for unimportant features to be removed, though not only the most important features are necessarily kept. Based on results of the data analysis, data could be considered more informative after processing, for example by combining them with other data. This is considered to be feature engineering, and this will be done in order to reach the best performing and efficient model.

After a number of iterations, a set of best performing models is reached. In order to understand these models' working, a more extensive feature importance analysis will be performed. These models are also assessed on their transferability. This transferability is tested by training on the dataset from one site, and testing on the dataset from the other site. This test will give insights into the practical application of the models, in order to understand the opportunities and limitations of the machine learning approach.

3.2.2. Neural Network

Artificial neural networks are computational models, which are designed to recognize patterns and solve complex problems. They process numerical data through interconnected neurons, ordered into layers. Between neurons, weights are given to the values, which are transformed within neurons through an activation function. The final layer is called the output layer, which is compared to a set of known parameter values, also known as target data. The error of the predicted values compared to the target values is measured by the loss function. Using backpropagation, weights are adjusted in order to minimize this loss function. Below, the implementation of these methods is presented.

Figure 3.1 schematically visualizes a neural network architecture. A typical neural network consists of layers of nodes. In any case, the neural network should consist of an input layer, hidden layer(s), and an output layer. The input layer receives the features used by the model. The hidden layer(s) receive the previous values, applies a linear weight to them, and then applies a non-linear activation function. This thesis uses the Scaled Exponential Linear Units (SeLU) activation function, which is shown in Figure 3.2. This activation function is particularly effective at preventing neurons from becoming stagnant, as both negative and positive values are possible as input without the output becoming zero (Klambauer et al., 2017). Therefore, it is favored over other activation functions. The output layer consists the values which are wished to be predicted.

The training process involves adjusting the weights between nodes to optimize predictions. This process typically includes:

1. **Forward Propagation:** Data that passes through the network, producing an output.
2. **Loss Calculation:** A loss function, which quantifies the error between the predicted and measured outputs, is used to assess the prediction.
3. **Backward Propagation:** Gradients of the loss function are computed with respect to each weight, and the weights are updated to minimize the loss.

This is performed during multiple iterations, or epochs. The loss value of the validation set is simultaneously calculated. When the validation loss does not decrease anymore for a defined number of epochs, training is stopped. In this thesis, training was stopped if validation loss did not decrease for 100 epochs. K-fold validation was performed in the training process of the models to ensure maximum use of the available data.

The chosen loss function for this problem is the mean squared loss (MSE). The Adaptive Moment Estimation (Adam) optimizer is used in the backward propagation. The scale at which each weight is

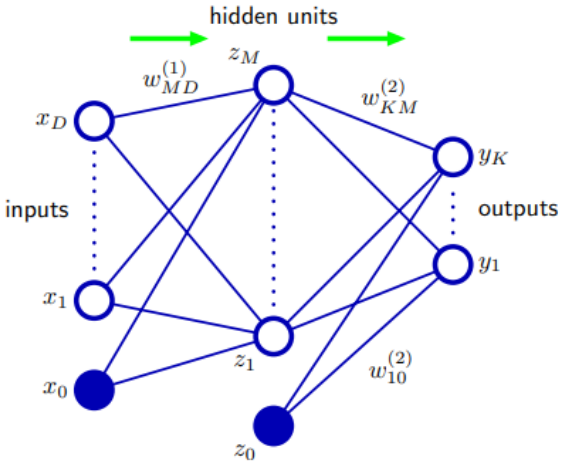


Figure 3.1: Schematic neural network architecture (Bishop, 2006).

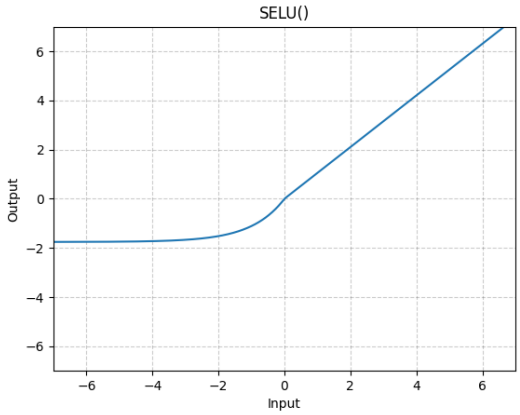


Figure 3.2: SELU activation function.

$$f(x) = \lambda * x \text{ if } x > 0$$
$$f(x) = \lambda * \alpha * (e^x - 1) \text{ if } x <= 0$$
where $\lambda \approx 1.0507$ and $\alpha \approx 1.67326$

changed is called the learning rate. In order to prevent overfitting, L2 regularization is performed. This ensures that test performance will not be significantly worse than training performance. L2 regularization essentially adds a penalty term to the loss function based on the squared values of the weights. The scale of this penalty term is determined by λ.

The number of hidden layers, size of the hidden layers, learning rate, and λ (regarding L2 regularization) are hyperparameters. They are parameters that are not part of the dataset or decided by the model during training, but they do influence the calibration of the model. These hyperparameters were tuned during the training process by minimizing the validation loss. The training/validation set was thus used for tuning the hyperparameters. This was done using GridSeachCV of the sklearn python package, with 5-fold cross validation and scoring based on mean squared error. For each hyperparameter, three possible values were considered, which are given in Table 3.1. More values could be considered, but increase computational cost significantly.

Hyperparameter			
Number of hidden layers	8	16	32
Hidden layer size	16	32	64
Learning rate	0.001	0.01	0.1
λ	1e-6	1e-7	1e-8

Table 3.1: Considered neural network hyperparameter values.

3.2.3. Random Forest

Random forests are a machine learning models that construct multiple decision trees during training and combines their outputs to make predictions (Breiman, 2001). A Random Forest model consists of the following:

- **Decision trees:** Individual models that recursively split the input data based on feature thresholds to form a tree structure.
- **Ensemble mechanism:** Combines the predictions of multiple decision trees to provide a final output, reducing the risk of overfitting and increasing generalization.

A schematic visualization of a random forest model is shown in Figure 3.3.

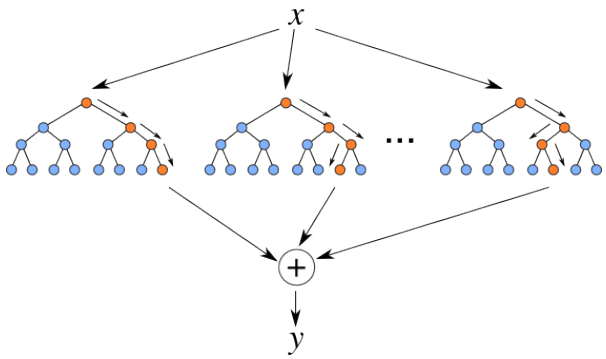


Figure 3.3: Schematic random forest architecture (Chaya, 2020)

The final prediction is the averaged value between the predictions of the decision trees. The training process of a random forest involves the following steps:

1. **Bootstrap sampling:** Random subsets of the training data are sampled with replacement to train each decision tree. This process, known as bagging, ensures diversity among the trees.
2. **Random feature selection:** At each split in a decision tree, a random subset of features is considered to determine the best split. This introduces additional randomness, improving the model's robustness.
3. **Tree growth:** Each decision tree is grown to its maximum depth without pruning, capturing detailed patterns in the data.

These steps collectively reduce variance and enhance the model's generalization capability Breiman, 1996. Bootstrap sampling also serves as an out-of-bag validation method, as explained earlier in Chapter 3.

The following hyperparameters primarily control the complexity, overfitting, and feature selection of random forest models:

- **Number of trees:** Specifies the total number of decision trees in the ensemble. A larger number generally improves performance but increases computational cost.
- **Maximum depth:** Limits the depth of each tree, controlling the model's complexity.
- **Minimum samples per split:** Determines the minimum number of samples required to split a node.

- **Number of features:** Defines the subset of features considered at each split. This parameter is crucial for balancing bias and variance.

Using the validation set, these hyperparameters were tuned to reach the highest possible model performance. Table 3.2 present the considered values of the hyperparameters. Within the sklearn package, which was used in this thesis, there are different options for the number of features considered per split. Three of them were considered, namely the square root, log squared and no limit options are considered.

Hyperparameter			
Number of trees	100	200	300
Maximum depth	10	20	None
Minimum samples per split	2	5	10
Number of features	square root	log2	none

Table 3.2: Considered random forest hyperparameter values.

3.2.4. SHAP Feature Importance

SHapley Additive exPlanations (SHAP) is a framework used for interpreting machine learning models. It quantifies the contribution of each feature to the prediction made by a model, resulting in insights regarding its decision-making process (Lundberg and Lee, 2017). SHAP stems from game theory, specifically the Shapley value concept, which attributes a payout to players in a game based on their contributions. In the context of machine learning, SHAP interprets features as players and the model's output as the game's payout (Shapley, 1953). Essentially, SHAP feature importance is a method of sensitivity analysis which can be applied to machine learning models.

The application of SHAP to machine learning models works as follows: the algorithm is passed the model and input training data. It then runs this data through a chosen explainer. In this thesis, the Permutation Explainer was used for neural networks and the Tree Explainer was used for the random forest algorithm. The explainers work differently, but both give a SHAP value to each data point. Below is an explanation on how these explainers work.

The permutation explainer works by estimating feature importance based on the impact of feature permutations on a model's predictions. For each feature, the explainer changes its values randomly while keeping other features fixed, and observes how the model's predictions change. By calculating the difference in predictions before and after the permutation, the method quantifies the contribution of the feature to the prediction. This process is repeated multiple times for each feature to ensure robust estimates, averaging the impacts over the different permutations. The resulting SHAP values represent the importance of each feature, providing insights into the model's decision-making process.

The tree explainer in SHAP (SHapley Additive exPlanations) is an approach specifically designed for tree-based machine learning models, such as random forests. Unlike the permutation explainer, the tree explainer makes use of the interpretability of the structure of the random forest to compute exact Shapley values for feature importance. It exploits the additive nature of the random forest and the splitting logic of decision paths to explain the model's output fairly regarding its input features. By passing through the tree, the explainer determines how much each feature contributes to the model's predictions at every decision point. The tree-specific approach ensures accurate and scalable explanations for the random forest models, resulting in an easy to understand feature importance.

In principle, the random forest already directly produces a feature importance. However, in order to keep a consistent result and assessment of the feature importance for the two different methods, it was chosen to make use of the SHAP analysis for both neural networks and random forests.

4

Data Overview and Analysis

This chapter aims to enlighten the reader on the available data, both quantitatively and qualitatively, which have been collected prior to the thesis project. Next to that, numerous analyses are performed, which explore the possibility of correlations found in the data. In this chapter, an overview of the test sites and acquired data is presented. This includes sensor data, test data, and weather data. The final section of this chapter discusses the considerations which have been made that influence how the data will be used in following steps.

4.1. Description of Test Sites

In situ data has been collected at two sites: the Maasdijk near Oijen and the IJsseldijk near Westervoort. Figure 4.1 shows the locations of the test sites. Both sites are located on primary dikes, thus representing key flood defence structures located throughout the Netherlands, as these are often built using similar materials and building regulations. The test site locations had specifically been selected in a place where the width of the relevant areas in the dikes' cross sections are sufficiently large to consider those areas homogeneous in lateral direction. These relevant areas differ per dike: for the Maasdijk, in situ measurements and tests originate from the crest of the dike, while for the IJsseldijk, data come mainly from the inner berm. Some tests were performed on the inner and outer toe of the IJsseldijk. However, since these were not paired with sensor data, only data from the inner berm are considered for the IJsseldijk.

Figure 4.1: Test site locations.

Test/sensor	Measurement period	Quantity
CPT	9/2019 - 12/2020	2x20 tests[1]
FVT	10/2019 - 9/2020	7 tests
Borehole	9/2019 - 8/2020	3 tests
Volumetric water content reflectometer	10/2019 - today	Every 10 minutes
Tensiometer	10/2019 - today	Every 10 minutes
Piezometer[2]	10/2019 - today	Every 10 minutes

Table 4.1: Performed in situ tests and placed sensors per site (same for both sites).

[1] CPTs were performed in pairs.
[2] Piezometers were only installed in the IJsseldijk.

Table 4.1 presents an overview of the in situ tests performed per site, including the period over which the measurements took place and the total number of tests performed, or in case of a sensor, the sensor's measurement frequency. CPTs which were always performed in pairs, i.e., two tests were performed during a test day. This was also the case for all but one of the FVTs. Tests were generally not spaced out equally in time; time between CPTs varies from 1 week to 5 weeks between October 2019 and December 2020. FVTs were performed thrice: October/November 2019, April 2020, and September 2020. Though time spacing was not consistent, it can be said that different saturation states of the dike were captured in the tests. Boreholes were performed in October 2019 and August 2020, and resulting soil samples were used for numerous laboratory tests. These laboratory tests were not used in this thesis and will thus not be described further. However, they are used to provide a description of the soil characteristics.

At each test site the sensors were placed at 4 depths, with depths ranging from 1 meter to 3.1 meters below ground surface. At the IJsseldijk, the sensors were placed at the inner berm of the dike, while at the Maasdijk, the sensors were placed on the crest. Only the in situ tests performed close to the sensor locations were used, although the exact distance between the test and the sensors was not recorded.

4.2. Weather Data

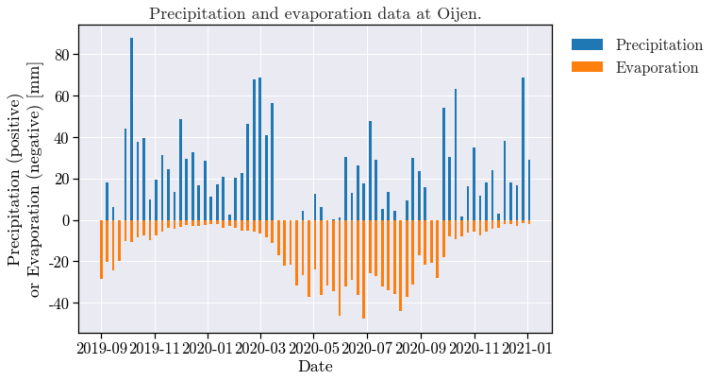
Since weather effects are considered to be the driving force for seasonal influence on dike slope stability, data from nearby weather stations are retrieved. This Section presents these weather data, and discusses whether these data could be used for data-driven machine learning models.

Figure 4.2 shows the weekly averages of the representative weather data, meaning precipitation and Makkink evaporation, for both test sites. This data was retrieved from the database of KNMI Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2025. Makkink evaporation combines different weather related parameters, such as radiation, temperature, and humidity.

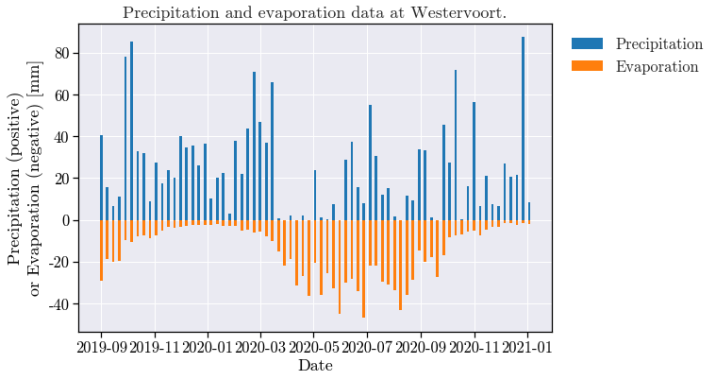
For the Maasdijk near Oijen, weather data were used from two KNMI measurement locations: Megen for precipitation and Volkel for evaporation. For the IJsseldijk near Westervoort, data were used from Deelen Airport.

It can be observed that generally speaking, precipitation and evaporation measurements are similar at the two test sites. Except for a 2-month long lower precipitation during April and May 2020, precipitation does not seem to have a season dependency. Evaporation, on the other hand, follows a clear sinusoidal pattern, where the least evaporation happens in the winter and the most during the summer.

The combination of these two parameters could be useful for machine learning models in order to have a better physical knowledge of the hydraulic state of the dike, which directly affects the mechanical state.



(a) Maasdijk, Oijen.

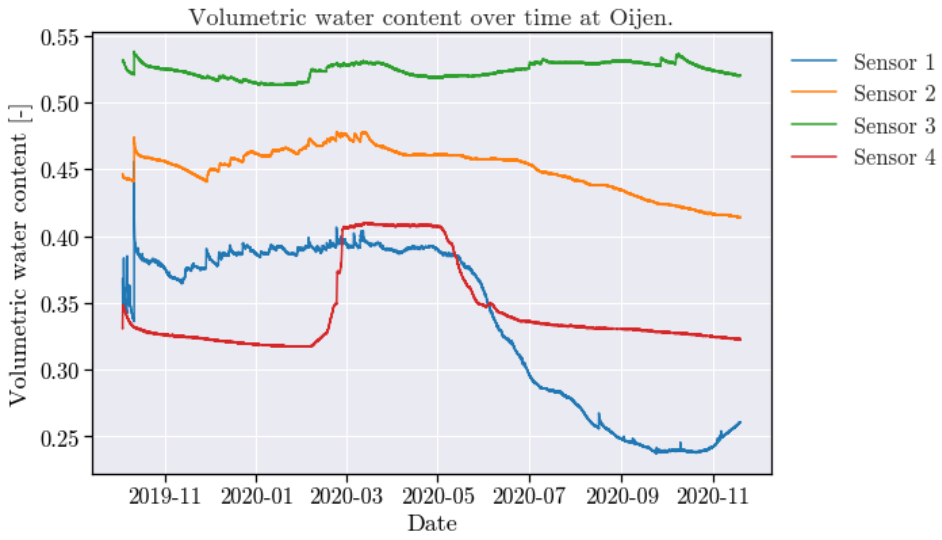


(b) IJsseldijk, Westervoort.

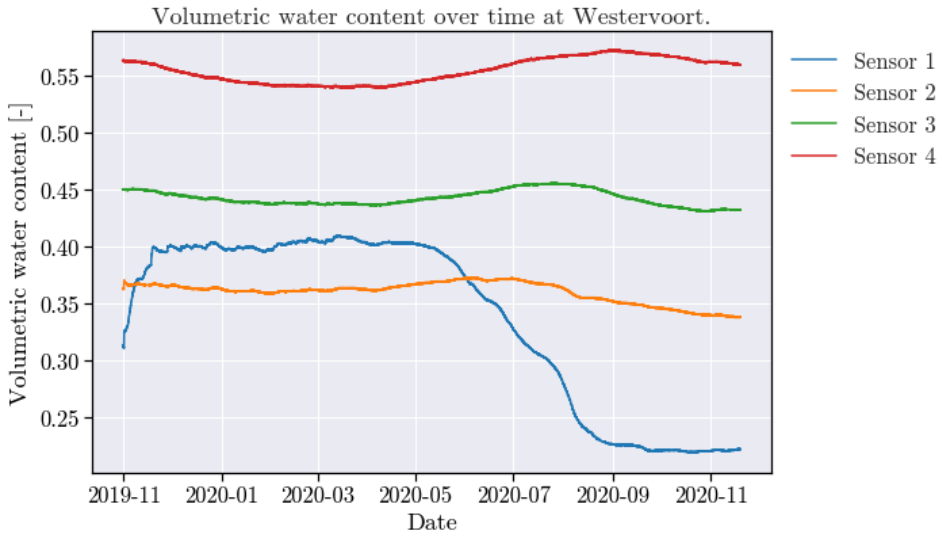
Figure 4.2: Weekly averages of the representative weather data at the two test sites.

4.3. Sensor Data

The recorded volumetric water content values and suction values are given in Figures 4.3 and 4.4 respectively. In Figure 4.3, it can be observed that from May 2020 - October 2020, there is a decrease in volumetric water content measured by sensor 1 at both test sites. At the Maasdijk, sensor 2 also measures a decrease in water content, but to a significantly lesser extent. At this test site, sensor 3's measurements stay relatively constant throughout the year. Sensor 4's measurements are at an increased level from March 2020 - June 2020, and remain at a relatively constant level otherwise. At the IJsseldijk, sensors 2, 3 and 4 all record relatively constant measurements throughout the year. However, one could observe a sinusoidal shape in the Figures, with a lag through the depth. To further elaborate: the water content measured by sensor 2 peaks around July 2020, measurements from sensor 3 peak around August 2020, and for sensor 4, it peaks around September 2020.



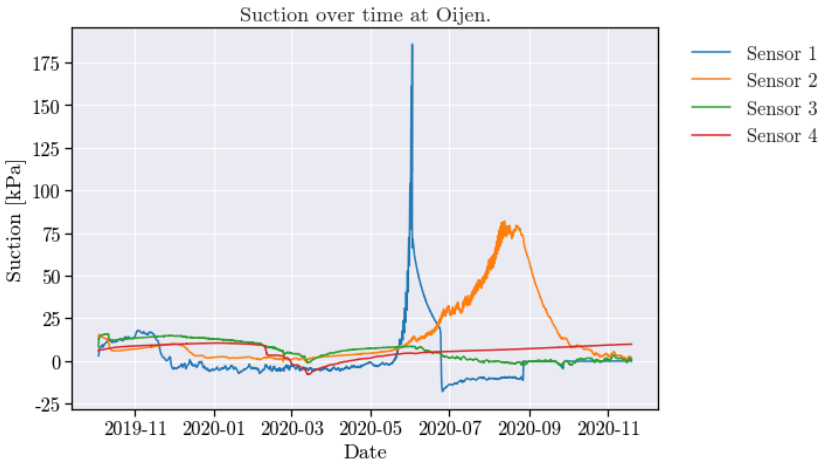
(a) Maasdijk, Oijen.



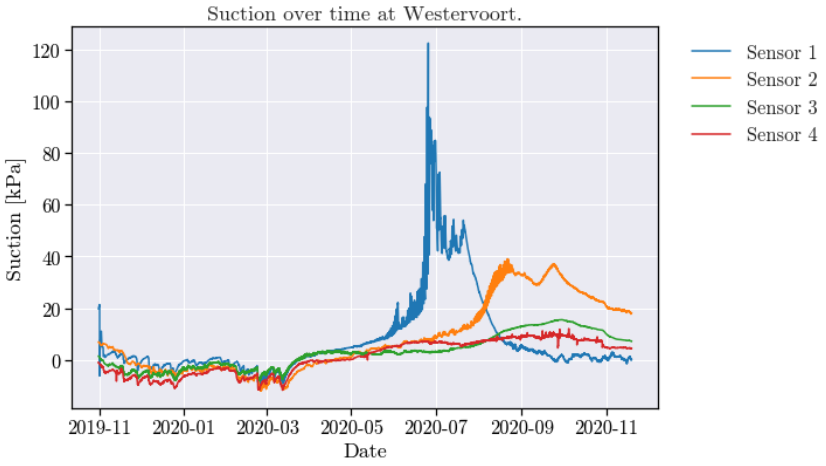
(b) IJsseldijk, Westervoort.

Figure 4.3: Volumetric water content measurements at the two test sites.

Figure 4.4 shows a large increment in suction measurements around June and July 2020 in sensor 1. As the measurements passed 100 kPa, this led to cavitation of the sensors, meaning that following measurements should not be considered reliable. Interestingly, a suction increase happens at a lower rate at the depth of sensor 2 at both test sites, with a slower and steadier increase of suction, and its peak around August - September 2020. Sensors 3 and 4 both measure relatively little variation of water content throughout the year, with increasingly smaller differences occurring at greater depth.



(a) Maasdijk, Oijen.



(b) IJsseldijk, Westervoort.

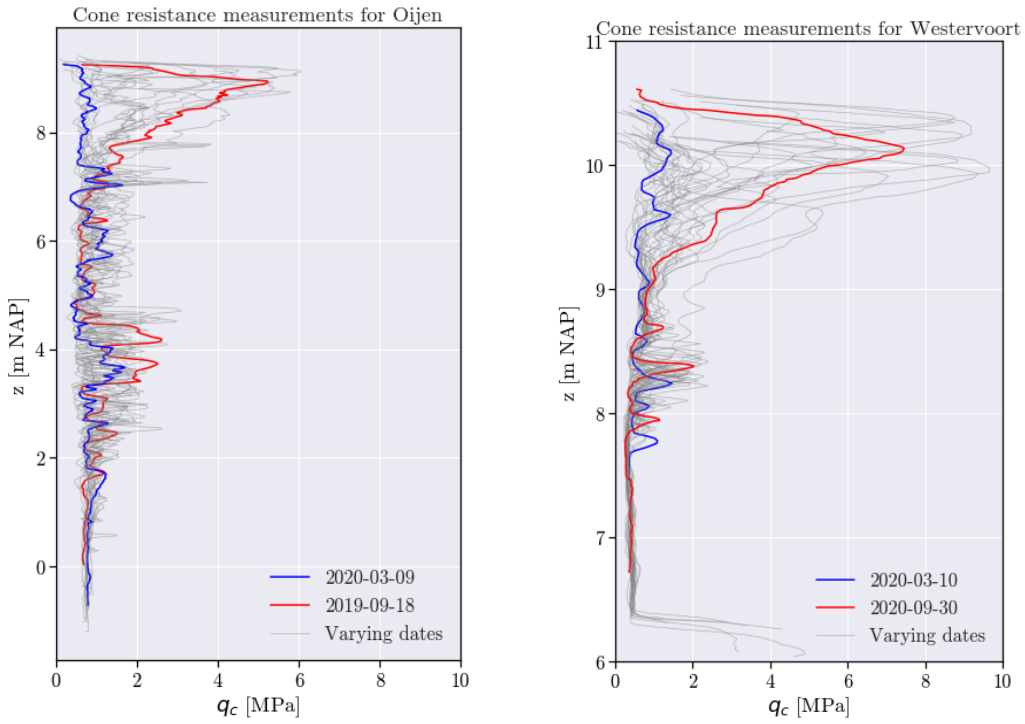
Figure 4.4: Tensiometer measurements at the two test sites.

4.4. Analysis of CPT Measurements

Figure 4.5 shows the cone resistance measurements of all tests performed for each location. Two tests have been chosen to be highlighted at each site, one for the wet season (blue) and one for the dry season (red), in order to show the difference in measurements between the two seasons. Other tests are only shown to give an idea of the CPT's behavior throughout the year. Evidently, cone resistance measurements at the same location change throughout the year in the initially unsatured layer. The size of this layer is estimated to be around 2 meters (7.3-9.3 m NAP) at the Maasdijk and around 1.5 meters (9-10.5 m NAP) at the IJsseldijk. In the initially unsaturated layer, the cone resistance ranges between 0.7 average lows to up to 6 MPa highs at the Maasdijk and between 0.8 average lows to up to 9 MPa peaks. Thus, in its dry state, the dike can exhibit cone resistance measurements that are 8.5 to 11.2 times higher than those in its wet state.

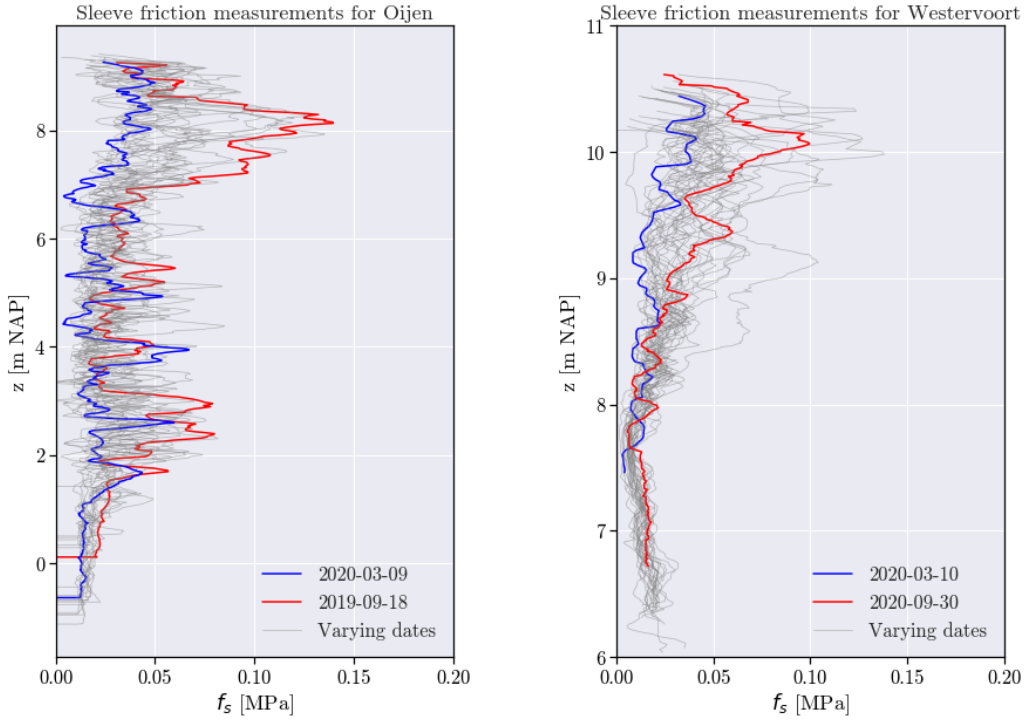
Figure 4.6 shows the same graphs, but now for the sleeve friction measurements. It is again observed that sleeve friction measurements in the dry season are higher than those in the wet season. The initially unsaturated layer appears to be of similar size as that of the cone resistance.

In the initially unsaturated layer, sleeve friction measurements range between 0.035 MPa average lows to 0.14 MPa peaks at the Maasdijk, and between 0.03 MPa average lows to 0.14 MPa peaks at the IJsseldijk. The multiplicative difference between sleeve friction measurements in the dry and wet state is thus up to 4-5 times. The effect of seasonal variance is thus smaller for sleeve friction measurements than for cone resistance measurements.



(a) Maasdijk, Oijen. **(b)** IJsseldijk, Westervoort.

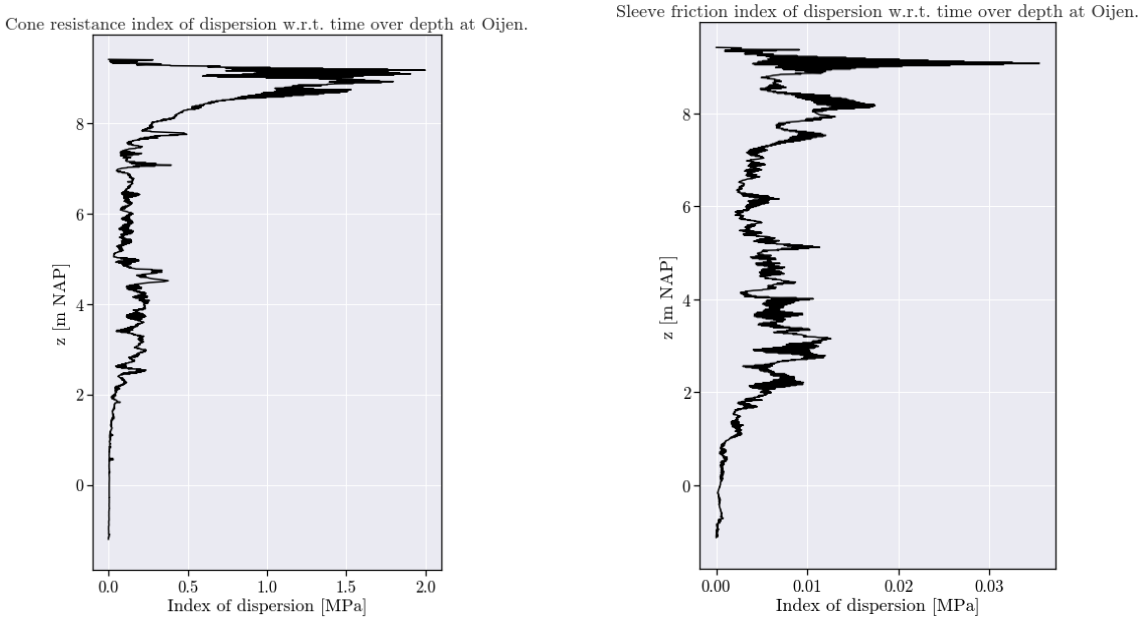
Figure 4.5: Cone resistance measurements at the two test sites.



(a) Maasdijk, Oijen. **(b)** IJsseldijk, Westervoort.

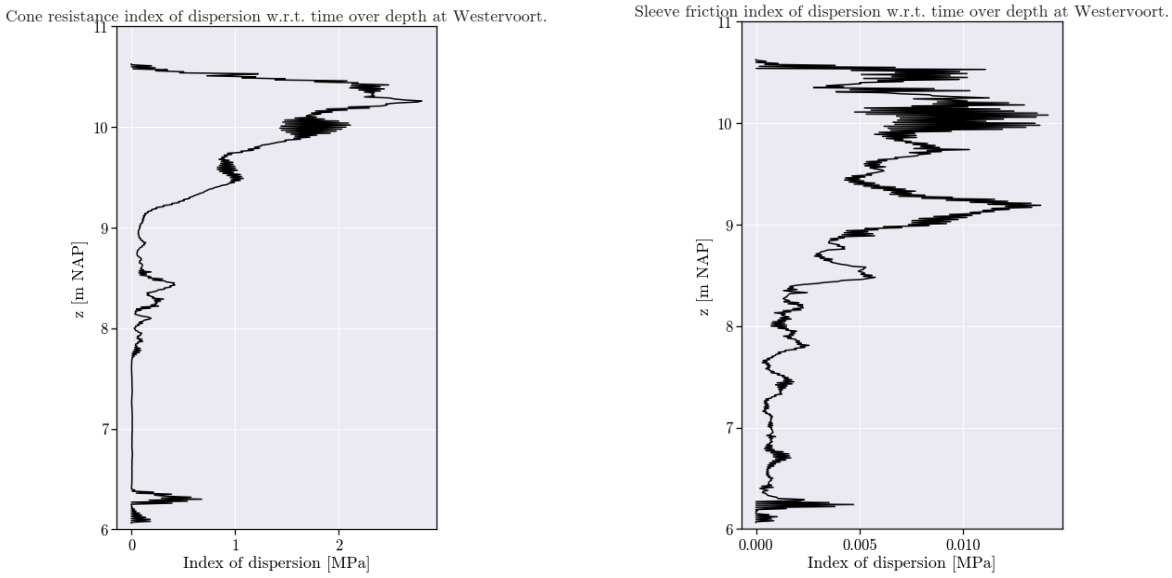
Figure 4.6: Sleeve friction measurements at the two test sites.

To further investigate the thickness of the initially unsaturated layer, the indices of dispersion (variance/mean) with respect to time of the cone resistance and the sleeve friction data are plotted over depth. This is shown in Figures 4.7 and 4.8. The index of dispersion essentially is a normalized variance, hence it is observed that the cone resistance measurements have a significantly larger index of dispersion than the sleeve friction measurements, thus confirming that the seasonal effect is noticed to a larger extent in the cone resistance measurements. It is also observed, however, that sleeve friction measurements may vary at depths where the cone resistance does not. For example, comparing Figures 4.8a and 4.8b, the index of dispersion is relatively large at around 9.2 m NAP for the sleeve friction measurements, but is nearing its minimum for the cone resistance measurements.



(a) Cone resistance. **(b)** Sleeve friction.

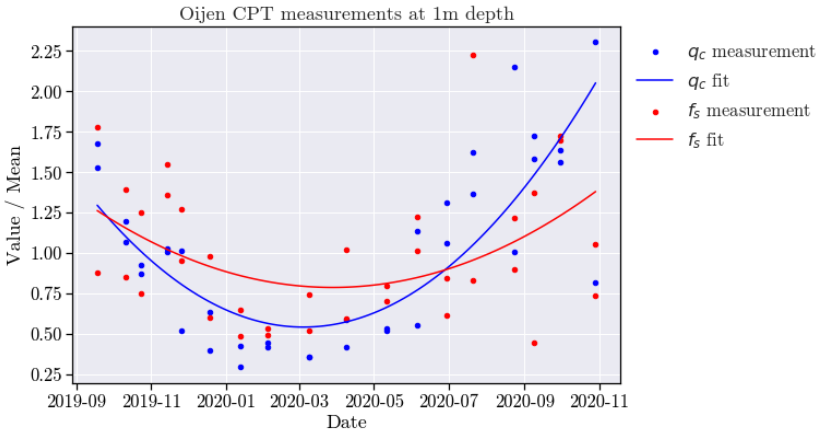
Figure 4.7: Index of dispersion w.r.t. time at Maasdijk, Oijen.



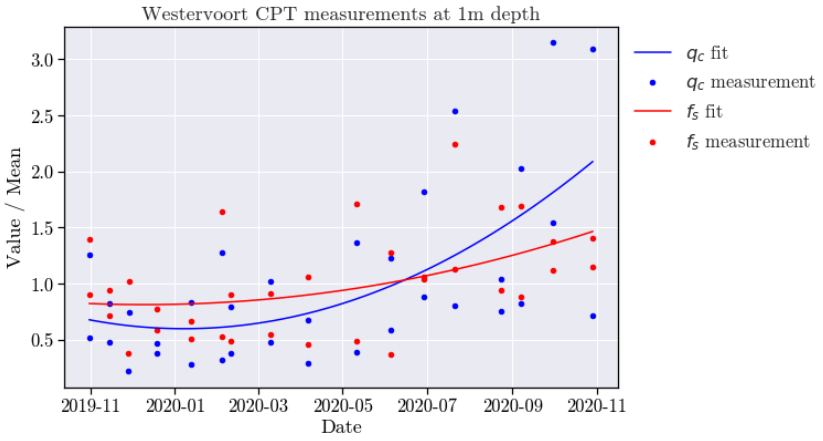
(a) Cone resistance. **(b)** Sleeve friction.

Figure 4.8: Index of dispersion w.r.t. time at IJsseldijk, Westervoort.

Looking into the seasonal behavior of the CPT measurements in the initially unsaturated layer, the cone resistance and sleeve friction measurements at 1 meter below surface are plotted throughout time in Figure 4.9. In this figure, a quadratic polynomial is fitted through the data, and values are scaled by the mean of all the values of its parameter. Again, it is observed that, under seasonal influence, cone resistance varies to a larger extent than sleeve friction.



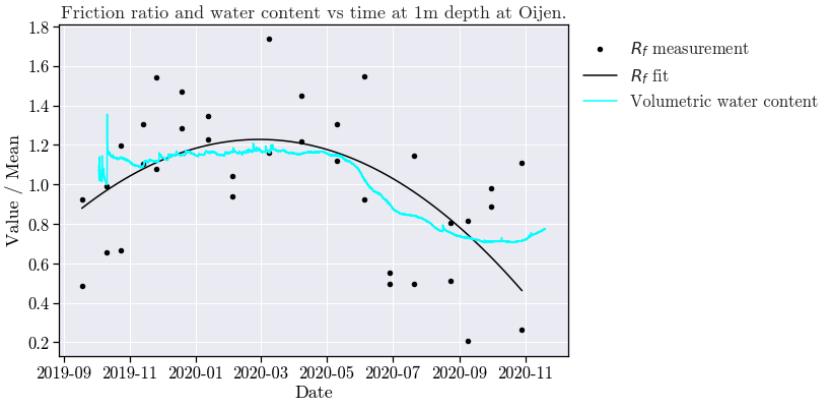
(a) Maasdijk, Oijen.



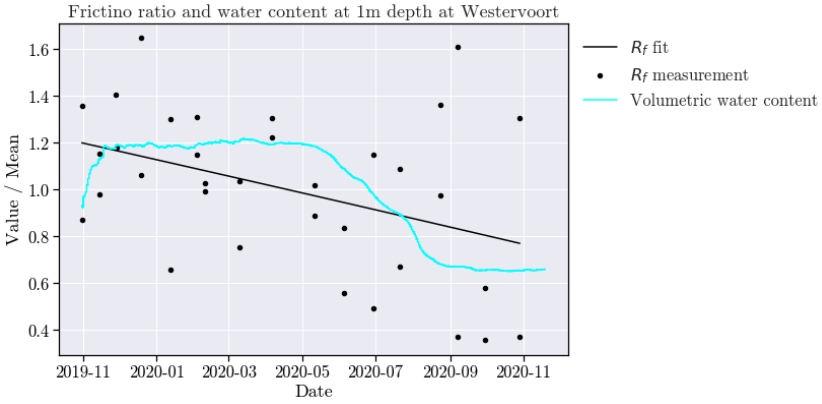
(b) IJsseldijk, Westervoort.

Figure 4.9: CPT measurements at 1m depth over time at the two test sites.

Since the conclusion can be made that cone resistance and friction ratio vary to a different extent under seasonal influence, the friction ratio too changes under the same seasonal influence. This is shown in Figure 4.10. In an attempt to relate the seasonally changing friction ratio, the volumetric water content by the sensor placed at the same depth (sensor 1 from Figure 4.3) is also shown. Both values are scaled by their mean. At the Maasdijk, a similar seasonal trend exists for the friction ratio and the volumetric water content. This is not observed at the IJsseldijk. In conclusion, a possibility could be considered for a data driven model which uses the friction ratio to assess the saturation of the dike.



(a) Maasdijk, Oijen.



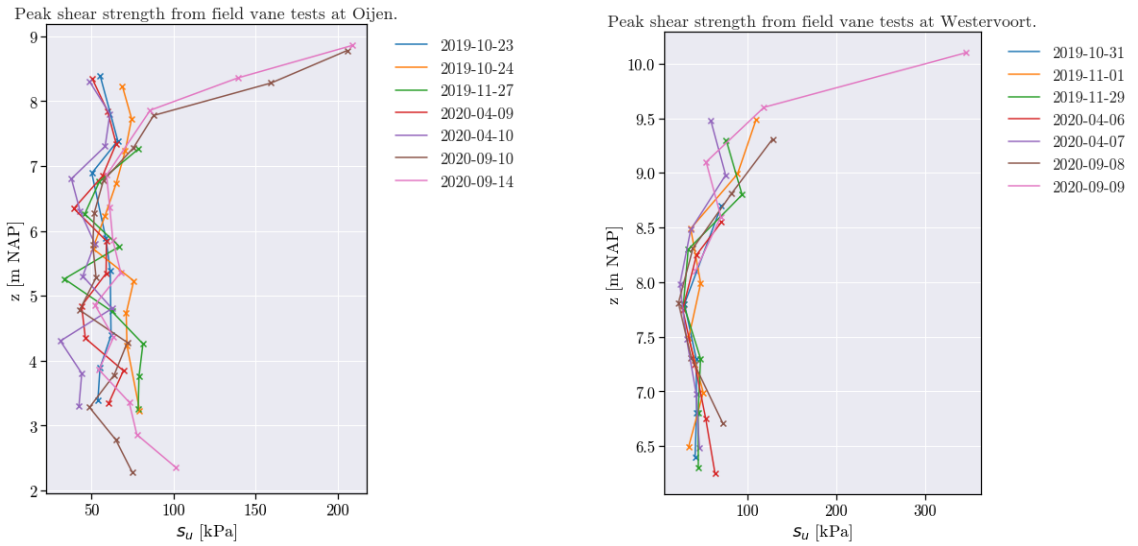
(b) IJsseldijk, Westervoort.

Figure 4.10: Friction ratio and water content 1m depth over time at the two test sites.

4.5. Analysis of FVT Measurements

Figure 4.11 shows the measured peak shear strength of all performed field vane tests at both locations. It is observed that the peak shear strength measured by field vane tests performed at the Maasdijk vary to a larger extent than those at the IJsseldijk. This could be caused by the fact that the tests were performed on the crest of the Maasdijk, while at the IJsseldijk, they were performed on the inner berm. At both locations, the peak shear strength varies to a larger extent the closer to the surface the measurement is, with high recordings of shear strength taking place in the dry season (September) and low recordings at other times. At depths over 1.5 meters below surface, no clear observation is made regarding seasonal effect on the measured shear strength.

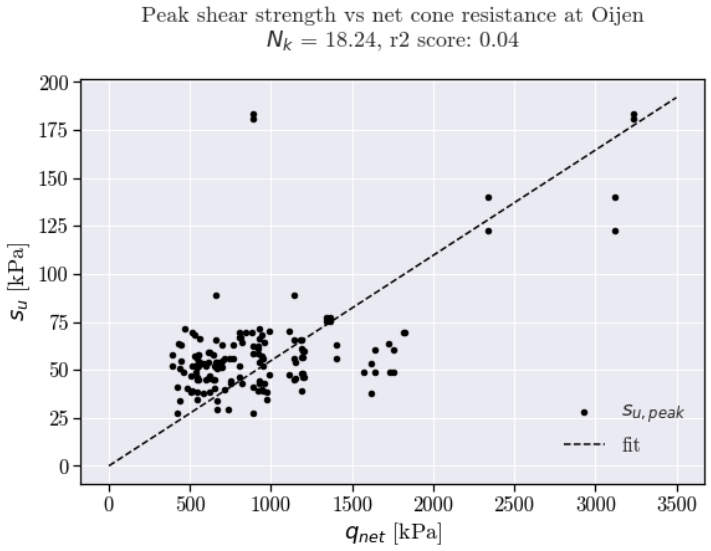
Combining the field vane test measurements with CPT measurements performed around the same date, Equation 1.1 can be considered to estimate N_k. Figure 4.12 presents this by plotting the peak shear strength against net cone resistance. N_k is estimated by fitting a linear relationship through the data. It is observed that different N_k is estimated at the two test sites, with a value of 18.24 at the Maasdijk and 14.08 at the IJsseldijk. The R^2 score between the measured peak shear strength and the predicted peak shear strength through Equation 1.1 is calculated in order to assess whether this shear strength prediction method is valid. These R^2 scores are mentioned in the Figure. At the Maasdijk, the R^2 score is near zero, meaning that no strong correlation exists between the actual and the predicted shear strength exists. At the IJsseldijk, the R^2 score is 0.71, meaning that there is a relatively good correlation. It is considered that this can be caused by the fact that shear strength varies to a lesser extent at the IJsseldijk, as shown by Figure 4.11, making its prediction through a simple method easier. It should then be considered that in areas of the dike where shear strength varies to a larger extent, a simple method might not suffice.



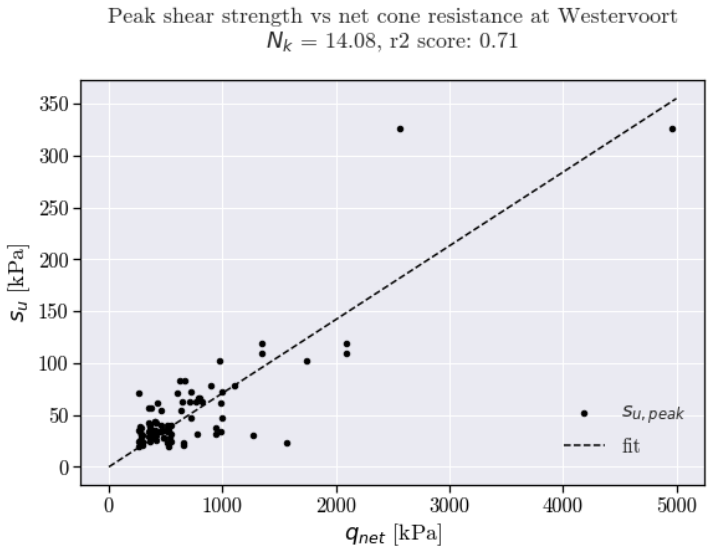
(a) Maasdijk, Oijen.

(b) IJsseldijk, Westervoort.

Figure 4.11: Field vane test results at the two test sites.



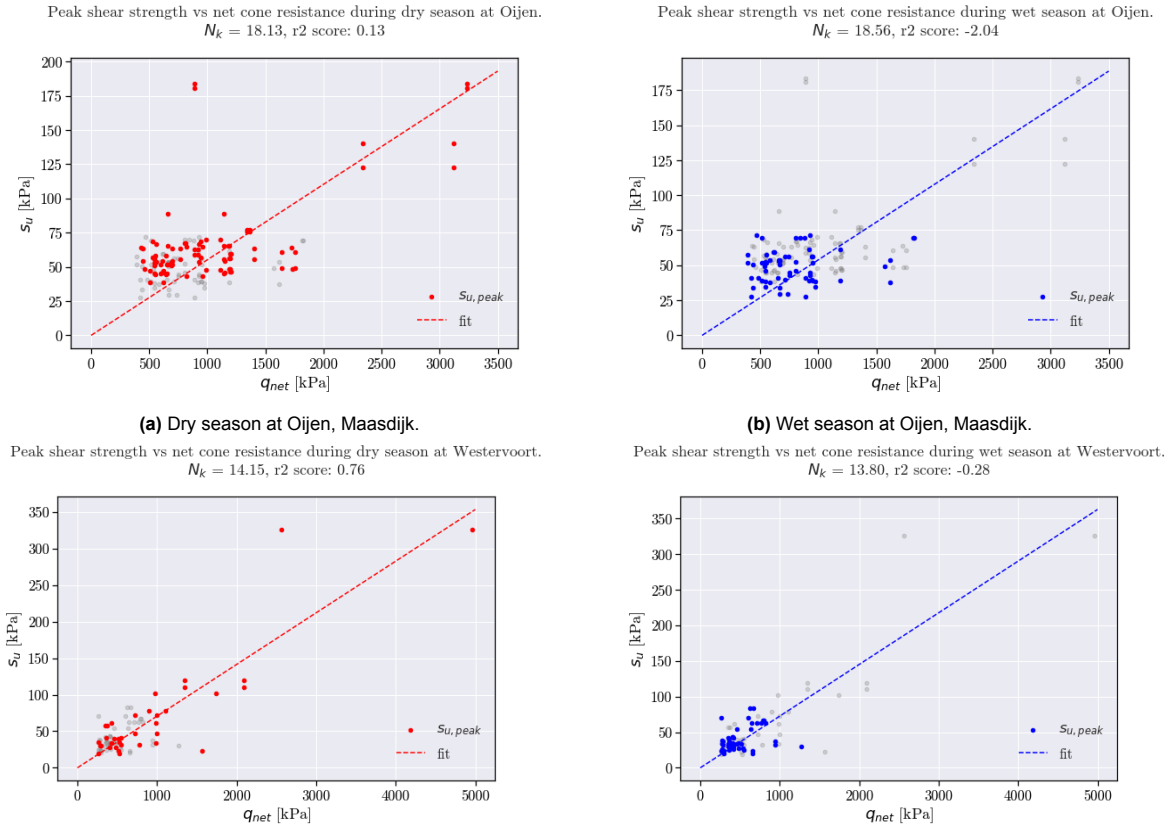
(a) Maasdijk, Oijen.



(b) IJsseldijk, Westervoort.

Figure 4.12: N_k approximation through Equation 1.1 for the two test sites.

In order to investigate whether seasonal effects are a large influence regarding the value of N_k, measurements are categorized by season, where the dry season runs from May-October and the wet season from November-April. The results after making this decision are plotted in Figure 4.13. Similar values for N_k are found, though the resulting R^2 scores differ largely between the dry and wet season, where the correlation is significantly better during the dry season than during the wet season. In practice, this is unfavorable, as the critical situation for slope stability would take place during the wet season.



(a) Dry season at Oijen, Maasdijk.

(b) Wet season at Oijen, Maasdijk.

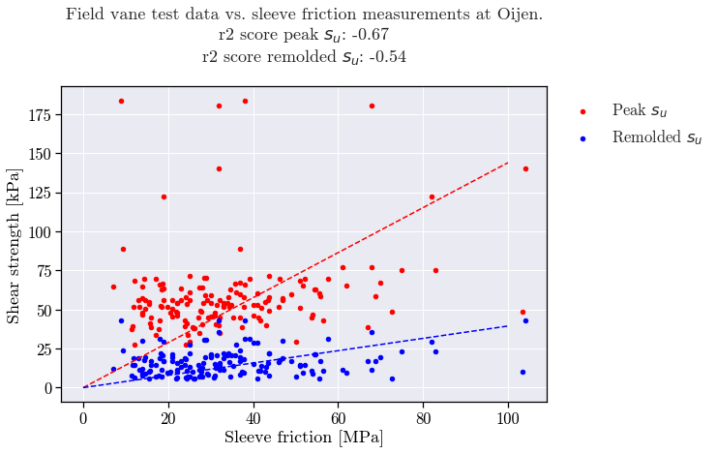
(c) Dry season at IJsseldijk, Westervoort.

(d) Wet season at IJsseldijk, Westervoort.

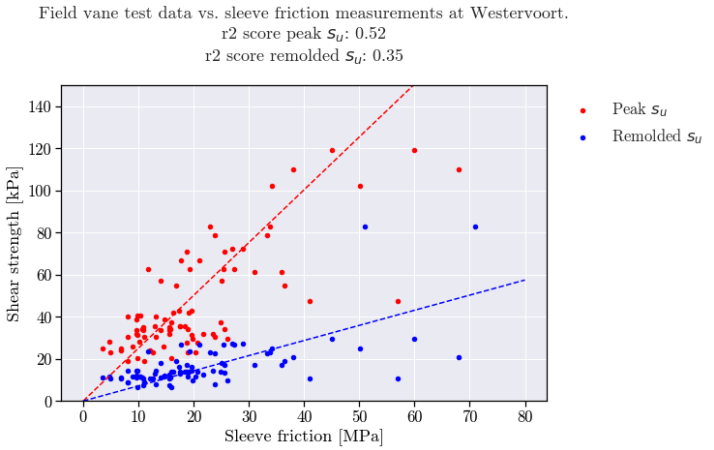
Figure 4.13: Field vane test results during dry and wet seasons at the two test sites: Oijen, Maasdijk and IJsseldijk, Westervoort.

Since no consistant correlation can be found between the peak shear strength and cone resistance, the possibility of a correlation existing between the shear strength and sleeve friction is explored. Results are shown in Figure 4.14. Next to the peak shear strength, the remolded shear strength is also given in this Figure. Robertson (2009) suggests a 1:1 correlation between sleeve friction and remolded shear strength, but this is not observed in the test data. The R^2 scores are again given for a linear correlation. The results show that at the Maasdijk, no clear correlation is observed, while at the IJsseldijk, a but positive, linear correlation can be seen, though the correlation is scattered.

To conclude, in trying to correlate the field vane test data to the cone penetration test data, no strong linear correlation was found between the critical, peak shear strength and both the cone resistance and the sleeve friction. However, there is reason to believe that a more complex analysis can exhibit correlation between the two tests because of two reasons. The first reason is that both the FVT results shown in Figure 4.11 and the CPT results shown in Figures 4.5 and 4.6 show that the measured parameters change under seasonal influence in the initially unsaturated layer. The second reason is that correlations between the CPT data and FVT data are found in the analysis of measurements from the IJsseldijk. They are possibly heavily scattered, however, and it appears that this is a more relevant correlation during the dry season than the wet season.



(a) Maasdijk, Oijen.



(b) IJsseldijk, Westervoort.

Figure 4.14: Field vane test results vs the measured sleeve friction at the two test sites.

4.6. Use of Data

In geotechnical context, a large dataset of repeated tests at a close location are available. However, in a machine learning context, the dataset can be considered small. When combining the data, the dataset for model type A (predicting θ) consists of 256 rows of data, and the dataset for model type B (predicting s_u) consists of 218 rows of data. Hence, in order to maximize the use of the available data, conscious decisions regarding the use of data should be made, which this section aims to provide.

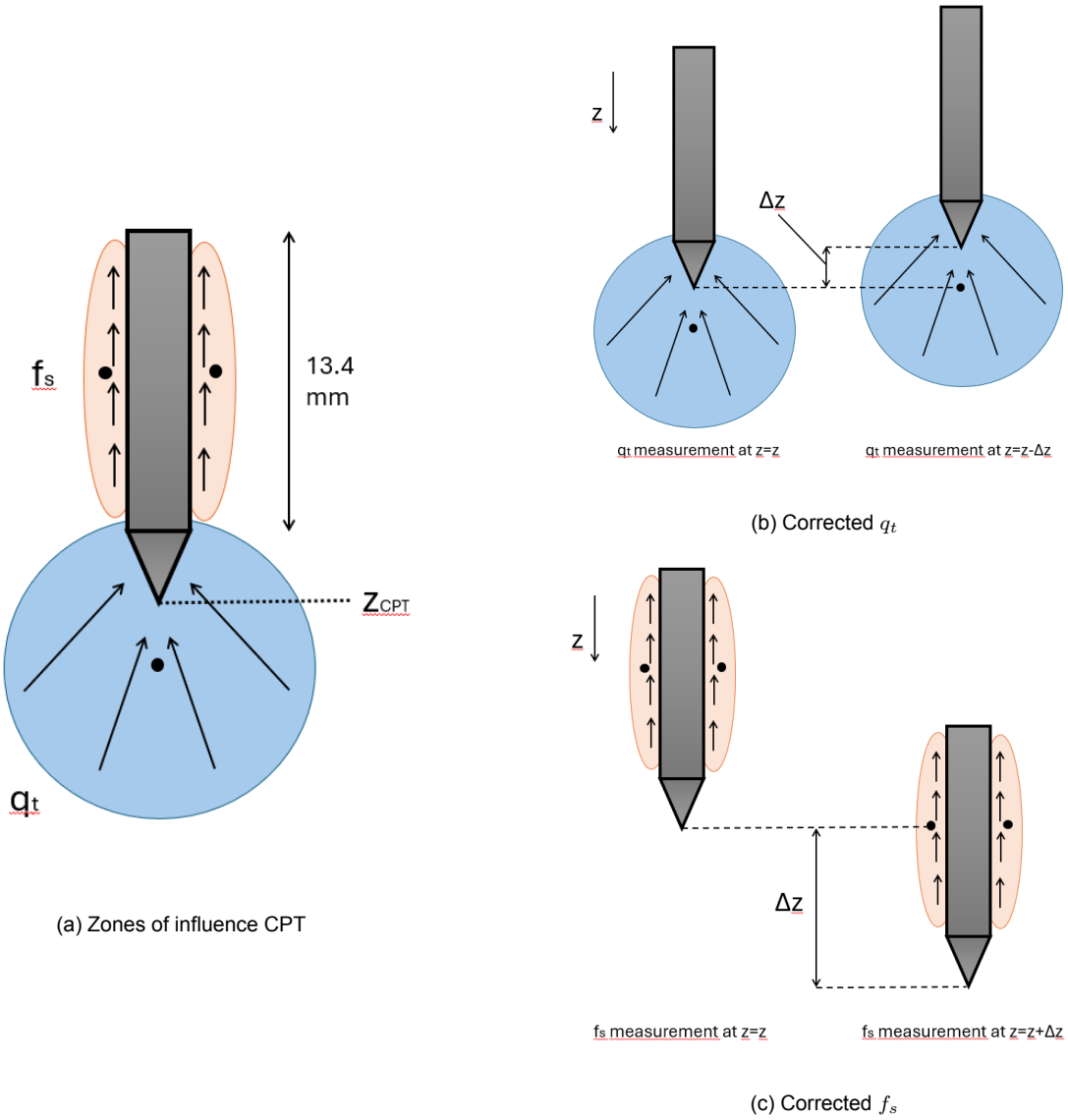
Firstly, taking into account the in-situ test frequency and the number of placed sensors, a discrepancy can be noticed between the discretization of in situ tests (CPTs, FVTs) and sensors with respect to space and time in the available dataset. The in situ tests are continuous in depth, but are relatively discrete in time compared to the sensors, which are continuous in time, but discrete in depth. Combining the FVT data with the sensor data is therefore not possible for the machine learning model, since this would lead to a very small dataset, which would be unusable for the models to train on. The result is that two different models are trained and tested: models that predict water content and models that predict shear strength.

It should be noted that, although data from the Maasdijk come from the crest of the dike, while data from the IJsseldijk come from the inner berm, they will be used in the same dataset initially in Chapter 5. The goal is to use the largest possible dataset for training, allowing the model to generalize effectively. Since the dataset is, in the context of machine learning, relatively small, the predictive performance might be limited.

It should be considered, that in reality, the assessor of a dike slope stability will likely not have sensor

data provided, nor are they likely to have field vane test data provided. Aiming for a practical use-case of the models, these measurements should thus not be considered as input data of a data-driven model, but rather as target data. Cone penetration tests are performed instead in the Netherlands, and weather data from nearby KNMI weather stations are publicly available. Therefore, using these data as input features would be practical.

However, it is important to note that it is yet unknown exactly how to use the CPT and weather data for the prediction of both shear strength and water content. For the CPT data, it should still be investigated from what depths, relative to the CPT tip depth, the most influential measurements are recorded. Figure 4.15 visualizes how this should be thought of. In this Figure, the difference between CPT tip and zone of influence is named Δz. It is decided that this Δz ranges somewhere between 0-10 centimeters in the relevant direction for q_t and f_s. This 10 centimeters is chosen based on CPT cone dimensions.



(a) Zones of influence CPT

(b) Corrected q_t

(c) Corrected f_s

Figure 4.15: CPT zone of influence correction by an unknown Δz.

For the weather data, this should be considered when selecting the representative period of the weather data, i.e. the time lag at which soil experiences weather effects. Both long-term and short-term weather data could be considered as relevant for the saturation state of a dike, thus also for the stability of the dike. Therefore, daily weather data is considered of the previous 4 weeks.

All in all, the following data can be used as raw input features, and the importance and usability of each

feature are yet to be understood as a result of using said feature in a machine learning model.

Raw features:

- Depth w.r.t. surface
- Net cone resistance ($q_n = q_c - \sigma_v$). Referring to the depth of the water content sensor or the FVT data, the following values are used:
  - CPT tip at the same depth
  - CPT tip 2cm above
  - CPT tip 4cm above
  - CPT tip 6cm above
  - CPT tip 8cm above
  - CPT tip 10cm above
- Sleeve friction (f_s). Referring to the depth of the water content sensor or the FVT data, the following values are used:
  - CPT tip at the same depth
  - CPT tip 2cm below
  - CPT tip 4cm below
  - CPT tip 6cm below
  - CPT tip 8cm below
  - CPT tip 10cm below
- Precipitation (P). Referring to the date of the relevant measurement (water content sensor/FVT), the following values are used:
  - On the same day
  - 1 day prior
  - 2 days prior
  - ...
  - 27 days prior
  - 28 days prior
- Evaporation (E). Referring to the date of the relevant measurement (water content sensor/FVT), the following values are used:
  - On the same day
  - 1 day prior
  - 2 days prior
  - ...
  - 27 days prior
  - 28 days prior

This gives a total of 71 possible raw input features. These raw features can be engineered, which could be helpful for two reasons. The first being the curse of dimensionality, entailing that a reduced size of the input layer could lead to increased performance. The second being that through knowledge-based engineering of features, information is given more efficiently, making it easier for models to identify patterns.

Below, the engineered features that are considered are mentioned.

Engineered features:

- Average net cone resistance over a depth interval ($q_{n,avg}$)

- Average sleeve friction over a depth interval ($f_{s,avg}$)
- Friction ratio (R_f or $R_{f,avg}$ if averaged)
- Average precipitation over the previous month (P_{month})
- Average evaporation over the previous month (E_{month})
- Short term weather parameter ($W = \frac{P_{week,avg} - E_{week,avg}}{z}$)

In the next Chapter, these features are used to build, train, and test machine learning models, as described in Chapter 3.

5

Results

This Chapter presents the test results of each constructed model, the feature importance analyses, and the results of the best performing models. Each Section discusses different input features and follows a modeling pipeline as discussed in Subsection 3.2.1. As two output types and two model types are considered, this means that four models are built for each section. Their naming is structured so that models named AX always predict θ and models named BX always predict s_u. In order to efficiently show results, numerous tables and figures can be found in the following appendices:

- Appendix A: Learning Curves
- Appendix B: Hyperparameters
- Appendix C: SHAP Summary Plots
- Appendix D: SHAP Dependence Plots

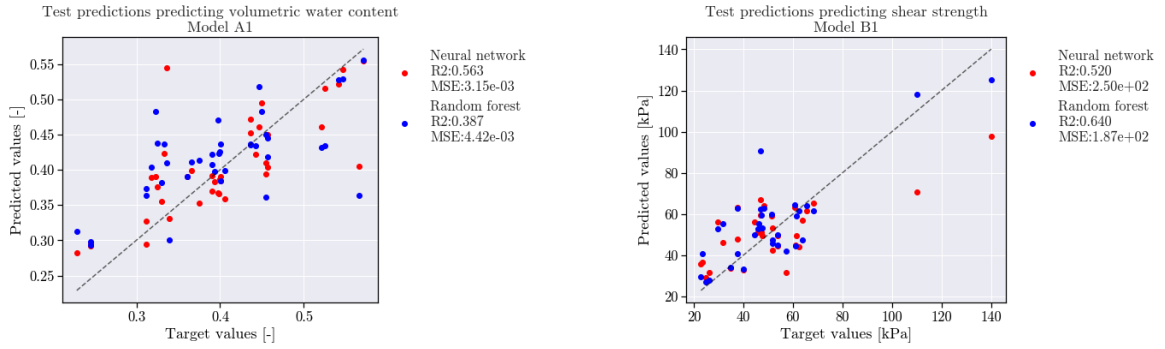
5.1. Raw Features

As a starting point, the models' inputs consist of the features in their raw state. This means, that the CPT measurements for the cone resistance and sleeve friction along the 10 centimeter depth interval around the depth of interest. For the weather data, daily precipitation and evaporation measurements of the previous 4 weeks are given. The depth and a bias term are also included in the input layer. For each model, the hyperparameters as tuned are given in Table B.1.

Figure 5.1 shows the test results, with target values on the X-axis and predicted values on the Y-axis. The test scores are also shown in this Figure. It is observed, that in the prediction of θ, the neural network greatly outperforms the random forest model. In the prediction of s_u, however, the random forest outperforms the neural network. It seems, that for very low values of s_u, the models tend to overpredict and for high values, the models tend to underpredict.

The summary plots showcasing the SHAP feature importance for each model are given in Figure C.1. These plots should be interpreted as follows: each dot represents a training data point. The color of the dot represents its value in accordance with the color bar on the right side of the plot. The larger the absolute SHAP value, the higher its importance for the predicted output. If the SHAP value is positive, that data point causes an increase of the predicted output, and when the SHAP value is negative, that data point causes a decrease of the predicted output. These SHAP values are unnormalized. A SHAP value close to zero indicates that there is little influence from that data point on the prediction. The figures thus show whether and how features influence predictions. It is important to note, however, that SHAP values are relative to a model's baseline prediction, and correlations between feature values and SHAP values do not necessarily indicate correlations between feature value and target value.

Figure C.1a shows that model A1-NN makes its predictions of θ using the sleeve friction measurements and the depth. How the sleeve friction measurements are used differs per Δz value along the depth interval. Some correlations are positive (f_s, $f_{s,+2}$), while others are negative ($f_{s,+6}$, $f_{s,+8}$, $f_{s,+10}$). Next



(a) Model A1.

(b) Model B1.

Figure 5.1: Test results models A1 and B1.

to that, the depth feature is important for its predictions, with a positive correlation. The influence of weather data can be considered negligible, according to the model.

Its random forest counterpart, model A1-RF, mostly uses depth, f_s, and q_t for its predictions, with importance in that order, as shown in Figure C.1b. A large depth dependence is observed, where low (i.e., shallow) depths lead to lower predictions of θ, medium depths lead to higher predictions of θ, and very high values are less important for predictions. The model suggests a negative correlation between the CPT parameters and the effect they have on the prediction. It is observed that f_s tends to become more important for the prediction of θ for very low feature values, while q_t becomes more important for very high values. The most important sleeve friction measurements are those measured at Δz equals +10 and +8 centimeters depth, while the most important net cone resistance measurements are those measured at Δz equal to 0 and -2 centimeters. Weather data is much less important in its predictions, though both long and short term precipitation and evaporation measurements appear to have a small influence on the model's predictions.

Results of the summary plot for model B1-NN show that predictions of s_u are made using sleeve friction measurements, and other features seem to be unimportant. The sleeve friction measurements increase in importance as the depth increases, with the most important measurements around Δz equal to +8 to +10 centimeters. A consistent positive correlation is observed, except for the sleeve friction at Δz = 0. However, this feature is less important for the model's predictions. The depth feature and weather features are significantly less important than f_s.

Similarly to model A1-RF, model B1-RF depends largely on depth, with very low values of depth leading to an increased predicted value of s_u. These are likely to be the few measurements measured at shallow depths in the dry season (see Figure 4.11), and no clear correlation between depth and s_u is actually observed by the model. Differently to model A1-RF, this time q_t is considered to be a more important feature than f_s. The most important measurements come from -6 and -8 centimeters Δz for q_t, and from +8 and +10 centimeters for f_s. The model again makes little use of the weather features provided in the input layer, although some weather features do influence the model to some extent. The correlations between feature value and SHAP value vary between positive and negative.

Conclusion

The main observations are as follows:

- When predicting θ, neural networks outperform random forests. When predicting s_u, random forests outperform neural networks.
- Neural networks make their predictions largely based on measurements f_s at different depths. For the prediction of θ inconsistent correlations between feature and SHAP value are found, while for the prediction of s_u, the correlations are mostly positive.
- Weather data is generally unimportant to all four models. Some measurements seem to slightly

influence predictions, and these come from different time deltas prior to the date of recording.

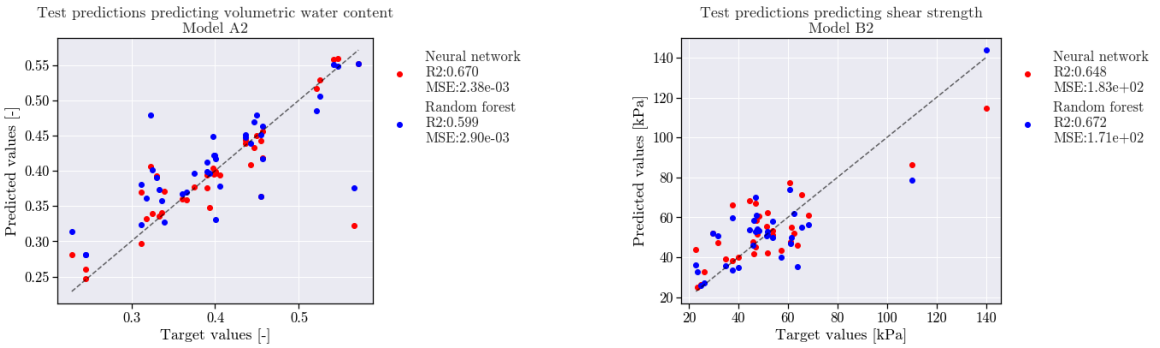
- According to random forest models, f_s is more important than q_t when predicting θ, while it is the other way around when predicting s_u. Also, depth is the most important feature, although no results show a clear correlation between depth and the according prediction.
- The most important measurements of f_s are measured at +8 to +10 centimeters Δz. The most important measurements of q_t are measured at -0 to -6 centimeters Δz.

Following the results, the decision is made to test the importance of the CPT features and weather features separately, of which the results are discussed in Sections 5.2 and 5.3, respectively. In this way, it can be observed whether weather data truly has little meaning to the model output, or if they were just overshadowed by the information that is already stored in the CPT data.

5.2. CPT-only Features

The input layers of the models discussed in this Section consist of the raw CPT measurements, depth, and a bias term (equal to 1). The hyperparameters are given in Table B.2.

The test results are shown in Figure 5.2. All models improved performance by removing weather features from the input layer. The largest increase in test performance was seen in model A2-RF. However, the neural network still outperformed the random forest for the prediction of θ, while the random forest slightly outperformed the neural network for the prediction of s_u.



(a) Models A2.

(b) Model B2.

Figure 5.2: Test results models A2 and B2.

Figure C.2 shows the summary plots of the SHAP feature importance. Mostly, the same observations can be made as those in Section 5.1. Although, by removing the weather data from the input, the neural network predicting s_u (model B2-NN) is finding a highly important, negative correlation between feature and SHAP value for f_s measured at $\Delta z = 0$. f_s measured at Δz = +4, +8 and +10 centimeters, however, still result in positive correlations between feature and SHAP value, with the most important depth at +10 centimeters Δz. Additionally, model B2-RF is ranking nearly all q_t features higher than f_s features, however depth is the most important feature, and some very low values for depth are associated with high SHAP values, which could possibly indicate overfitting.

Conclusion

The main observations are:

- By removing raw weather features, test performance for all models increased.
- The SHAP analysis found features to correlate similarly to the model's predictions with and without weather features.
- Along the depth interval, some CPT features became more important after removing weather features.

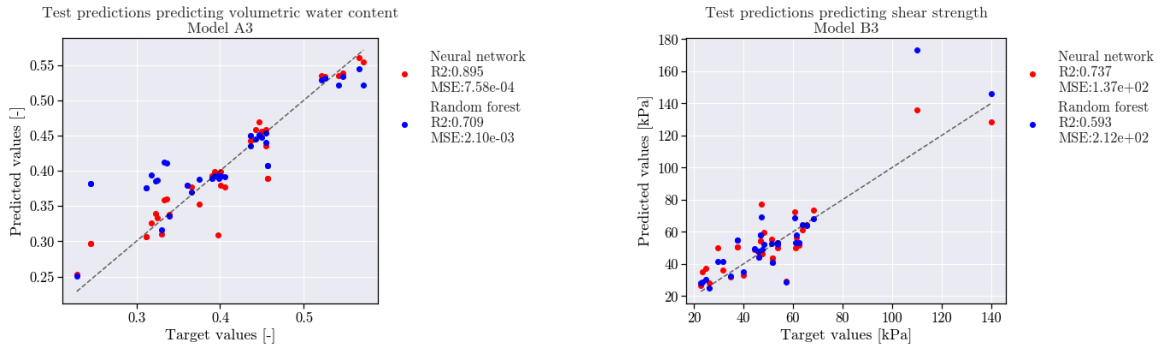
- Depth was not an important feature for the neural networks, but it was the most important for the random forests, with SHAP values indicating the possibility of overfitting occurring.

The improvement of model performance could suggest that the models in Section 5.1 are suffering from the curse of dimensionality, thus it is preferable to reduce the size of the input layer. Additionally, it could mean that enough information is captured within the CPT measurements to be able to bypass the weather data as features.

5.3. Weather-only Features

The input layer for the models discussed in this Section are the raw weather features, depth, and a bias term. Hyperparameters are again given, in Table B.3.

The test results are presented in Figure 5.3. The test scores record a significant increase in performance when predicting θ for both the neural network and the random forest by removing the CPT data. Especially high values are accurately predicted. The performance of the neural network predicting s_u also improved, while the performance of the random forest decreased.



(a) Model A3.

(b) Model B3.

Figure 5.3: Test results models A3 and B3.

By looking at the results of the SHAP feature importance shown in Figure C.3, the test results can be explained. It is observed that all models are basing its predictions largely on the depth. When predicting θ, this results in high test performance, but it can be concluded that the models are overfitted on the depth feature. Model B3-NN was able to find a precise correlation between feature and SHAP value of the depth, which was found to a less precise extent by model B3-RF.

Conclusion

The main observations are:

- By removing CPT data from the input, test results for the prediction of θ improve, while it has inconsistent consequences for the prediction of s_u.
- SHAP analysis shows that this conclusion is deceiving, as predictions are largely based on depth. These models are thus not generalizing well, but are rather predicting based on the scarce amount of depth locations of the sensors and FVTs to predict.
- Weather data was still unimportant for models.

Although the experiment to remove CPT features did not result in a change of the importance of weather features themselves, it did result in the important observation that there is a large risk of overfitting on depth, especially for models predicting θ. Next to that, the fact that weather features are still unimportant to models demonstrates that the weather features are unable to be useful in their raw state.

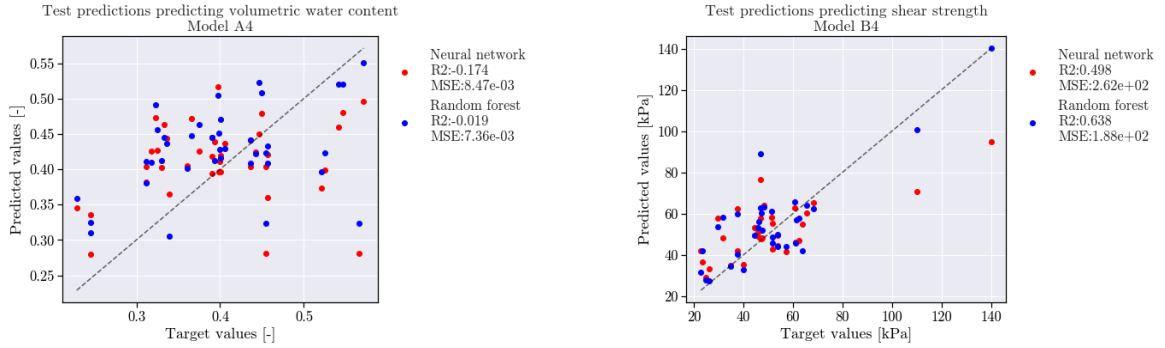
To mitigate the risk of overfitting, subsequent models will not make use of depth as an input feature anymore. For models predicting θ, it is observed that making depth-based predictions would result in

the best possible test predictions, but this would not result in a robust model. For models predicting s_u, overfitting could also occur to an unknown extent.

5.4. Raw Features Without Depth

To acquire a new baseline, the model inputs as used in Section 5.1 are again used, but this time without depth as a feature. Hyperparameters are given in Table B.4.

The test results are shown in Figure 5.4. It is observed that no accurate predictions of θ can be made by both the neural network and the random forest, as the R^2 score reduces to negative values for both models. The performance drop of model A4-NN compared to A1-NN is especially interesting, as depth was only the fifth most important feature for model A1-NN. Its removal, however, caused a decrease in R^2 score from 0.563 to -0.174. For the prediction of s_u, the neural network loses performance compared to model B1-NN. A relatively small performance loss is also seen for the random forest.



(a) Model A4.

(b) Model B4.

Figure 5.4: Test results models A4 and B4.

The SHAP summary plots are given in Figure C.4. A consistent negative correlation is observed for f_s between the feature and SHAP value. The f_s features are ranked by importance in accordance to their Δz value, with $\Delta z = +10$ as most important and $\Delta z = 0$ as least important. The model does not find importance in weather features.

Model A4-RF shows similar feature importance results as those of model A1-RF. Again, it is observed that q_t becomes more important to the model when the feature value is high, while f_s becomes more important when the feature value is low. Additionally, model A4-RF finds some value in the weather data from different time deltas.

Models B4-NN and B4-RF show similar results as before, with the neural network basing its predictions solely on f_s, and the random forest combining q_t and f_s to improve the performance over that of the neural network. No clear correlation is yet found between weather features and s_u by the random forest, but the features are used to a small extent.

Conclusion

The main observations are as follows.

- The predictions of θ suffer greatly from the removal of depth from the input layer.
- The predictions of s_u suffer slightly from the removal of depth from the input layer.
- Weather data is still unable to be of important to the models, though the models do attempt to use them.

Overall, the test performance of the models suffered by removing the depth-bias. As it is expected that there is an important correlation between weather effects and θ, the subsequent models should attempt to incorporate weather data through engineered features, as described in Section 4.6. If this improves

the prediction of θ, it is expected to also improve the prediction of s_u, as these two output parameters are themselves also expected to be correlated.

5.5. Engineered Weather Features

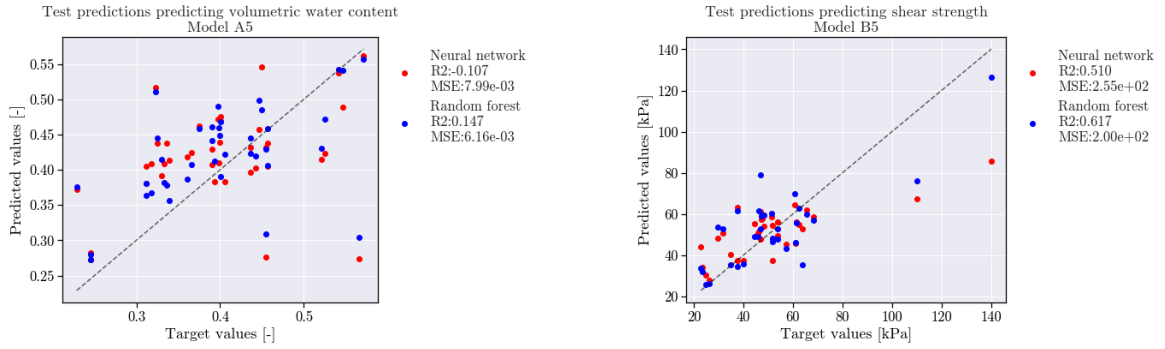
To attempt to help the models find information in weather features, these features are no longer used in their raw state, but instead in a feature-engineered state, as described in Section 4.6. They are:

- Average precipitation over a long period (>1 week) (P_{long})
- Average evaporation over a long period (>1 week) (E_{long})
- Short term weather parameter ($W = \frac{P_{week,avg} - E_{week,avg}}{z}$)

The first two are meant to capture the long term saturation state of the dike. The last feature is meant to capture the short term effects experienced by the dike under weather effects. The parameters is scaled by the depth, as weather effects are expected to decrease at deeper depths.

The hyperparameters as tuned are given in Table B.5.

The test results are given in Figure 5.5. It is observed that most models increase in performance by using engineering weather features. Only model B5-RF suffered slightly in performance from the change. Although the models improved in performance, the predictions of θ were still not accurate. Both in the prediction of θ and s_u, the random forest models outperform the neural networks.



(a) Model A5.

(b) Model B5.

Figure 5.5: Test results models A5 and B5.

The feature importance results are shown in Figure C.5. Once again, random forest models are able to make use of more different features, while neural networks base their predictions almost solely based on f_s. Model A5-NN finds inconsistent correlations between feature and SHAP value of f_s, as was previously the case for models A1-NN, A2-NN and A3-NN. Model A5-RF is able to use the engineered weather features, but no clear correlations are found.

In the prediction of s_u, the random forest model B5-RF makes use of W, P_{month} and E_{month}. Correlations between feature and SHAP value are not yet very clear, although it does seem that a positive correlation is found for P_{month}, and negative for E_{month}, which is a counterintuitive observation.

Conclusion

The main observations are:

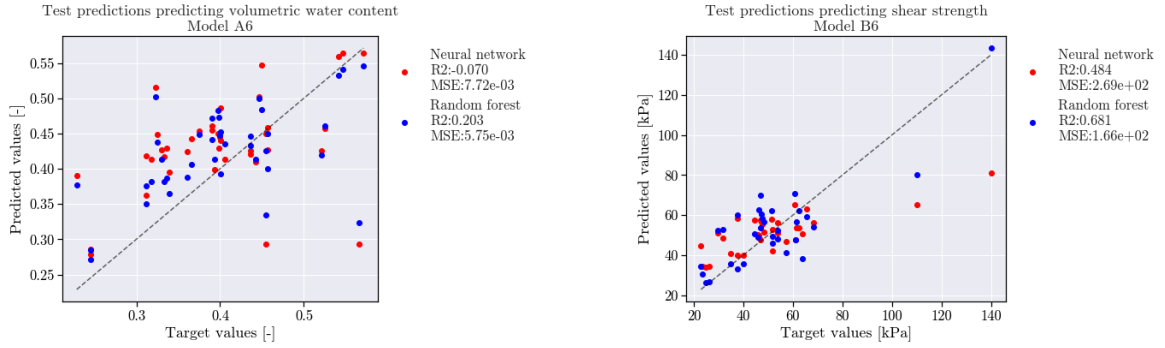
- Although the models use the engineered weather features to a different extent, it is observed that test performance improved for most models improved by swapping raw weather features for engineered ones.
- Model B5-RF finds counterintuitive correlations, which could be the reason why the model performance decreased.

As feature engineering of weather features generally resulted in an increase of performance, it could be the case that the raw weather features were causing the curse of dimensionality to occur. However, the engineered weather features do not yet help the θ-predicting models to such an extent that accurate predictions are achieved. Hence, subsequent models should consider adding and/or engineering features further to improve the information stored in the input layer.

5.6. Engineered Weather Features and Friction Ratio

As demonstrated in Figure 4.10, the friction ratio could be a meaningful feature to predict water content. Hence, the friction ratio is introduced as a feature in addition to the features of Section 5.5. The hyperparameters are given in Table B.6.

Test results are shown in Figure 5.6. Except for model B6-NN, an improvement of the test performance is observed. Still, no accurate predictions are made for θ.



(a) Model A6.

(b) Model B6.

Figure 5.6: Test results models A6 and B6.

The summary plots in Figure C.6 show that R_f is mainly an important parameter to the neural network models, and less important for random forest models. Although q_t is not being used by these neural network models, some of its information is eventually passed to the model through R_f. As a result of adding R_f, model A6-NN performs better than model A5-NN, although correlations are inconsistent with each other and with the physical expectations. No improvement of performance was observed going from model B5-NN to B6-NN, although model B6-NN seems to find meaning in R_f.

The random forest models rank R_f among the least important parameters. Even still, an improvement of the performance is observed.

Conclusion

The main observations are:

- In general, the addition of the friction ratio caused an improvement in performance.
- No consistent correlation for R_f was found by the models.

As has been observed through the SHAP feature analysis of the neural network models so far, difficulty may occur in the interpretation of raw CPT parameters, where multiple input features are expected to, in some extent, result in the same correlation. This was most often not the case however. In subsequent models, it should be explored how to simplify CPT input features for the models, in order to promote consistent interpretation.

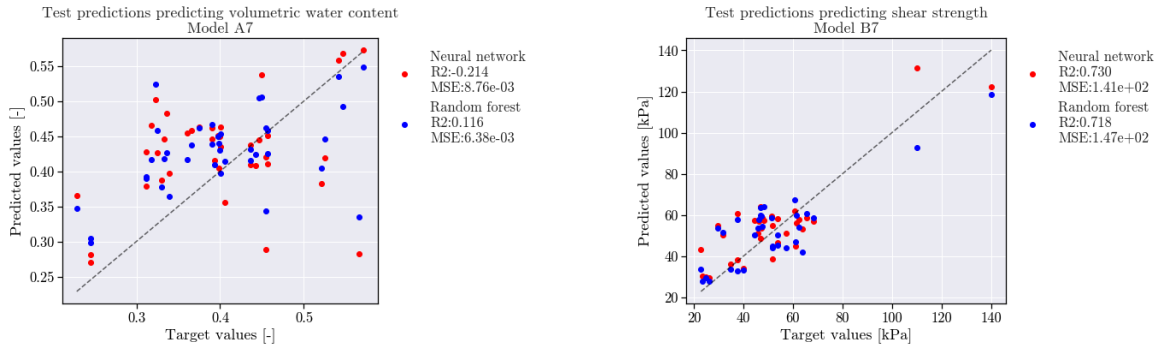
5.7. All Engineered Features

In this Section, all of the features are engineered to attempt to simplify the input layer, which should reduce the occurence of the curse of dimensionality, and make interpretation of features easier for the models. All of the engineered features mentioned in Section 4.6 are used, which are:

- Average net cone resistance over a depth interval ($q_{n,avg}$)
- Average sleeve friction over a depth interval ($f_{s,avg}$)
- Average friction ratio over a depth interval ($R_{f,avg}$)
- Average precipitation over a long period (>1 week) (P_{long})
- Average evaporation over a long period (>1 week) (E_{long})
- Short term weather parameter ($W = \frac{P_{week,avg} - E_{week,avg}}{z}$)

The hyperparameters are shown in Table B.7.

The test results are shown in Figure 5.7. A performance decrease is observed for model A7-NN and A7-RF by using the engineered CPT features rather than the raw ones. A performance increase is, however, observed for models B7-NN and B7-RF. Furthermore, it is noted that the models do not consistently overestimate or underestimate s_u.



(a) Model A7.

(b) Model B7.

Figure 5.7: Test results models A7 and B7.

Summary plots of the SHAP feature importance analysis are shown in Figure C.7. Models A7-NN and B7-NN both consider f_s to be the most important parameter, followed by R_f. The rest of the parameters are not important to the models predictions. A negative correlation is found for f_s and a positive for R_f when predicting θ, and vice-versa when predicting s_u. This is in line with expectations. Both model A7-RF and model B7-RF are able to use all features, and correlations are in line with expectations.

Conclusion

The main observations are:

- Models predicting θ suffer from the engineering of CPT features.
- Models predicting s_u improve following the engineering of CPT features.
- Neural networks are able to make predictions based on f_s and R_f alone, while random forests make use of all given features. With the engineered features, however, the neural network outperforms the random forest in the prediction of s_u for the first time.

5.8. Feature Interpretation by Best Performing Models

All features mentioned in Section 4.6 have been explored. Thus, this Section will not change the features, but rather evaluate how features are interpreted and used by the models that recorded the best performance.

As Section 5.3 concluded that the use of depth as a direct feature could introduce an overshadowing bias, thus increasing the risk of overfitting on the dataset, no models that use this feature are considered further. The reason for this is, that a generalized model is wished for, and an overfitted model would not comply.

All model test performance is summarized in Table 5.1.

Model	R^2 score	MSE
A1-NN	0.563	0.00315
A2-NN	0.670	0.00238
A3-NN	0.895	0.000758
A4-NN	-0.174	0.00847
A5-NN	-0.107	0.00799
A6-NN	-0.070	0.00772
A7-NN	-0.214	0.00876

(a) Neural networks type A.

Model	R^2 score	MSE
A1-RF	0.387	0.00442
A2-RF	0.599	0.00290
A3-RF	0.709	0.00210
A4-RF	-0.019	0.00736
A5-RF	0.147	0.00616
A6-RF	0.203	0.00575
A7-RF	0.116	0.00638

(b) Random forests type A.

Model	R^2 score	MSE
B1-NN	0.520	250
B2-NN	0.648	183
B3-NN	0.737	137
B4-NN	0.498	262
B5-NN	0.510	255
B6-NN	0.484	269
B7-NN	0.730	141

(c) Neural networks type B.

Model	R^2 score	MSE
B1-RF	0.640	187
B2-RF	0.672	171
B3-RF	0.593	212
B4-RF	0.638	188
B5-RF	0.617	200
B6-RF	0.681	166
B7-RF	0.718	147

(d) Random forests type B.

Table 5.1: Test performance of all models.

Of those that did not use depth as a feature, the best performing A-type models were models A6-NN and A6-RF, while the best performing B-type models were models B7-NN and B7-RF. However, since A6-NN and A6-RF had poor test predictions with low R^2 scores, they are not analyzed further, but instead A2-NN and A2-RF are. This is done in order to investigate and highlight how depth was interpreted by these models. These and models B7-NN and B7-RF are used to evaluate feature interactions deeper using the SHAP analysis.

To understand the correlation between feature value and SHAP value, so-called dependence plots can be made. The summary plots previously discussed in this Chapter already give an indication whether a correlation is positive or negative through the colors of data points, but non-linear correlations are difficult to notice. Next to dependence plots, the tree explainer used for random forest models can be used to calculate SHAP interaction values, which quantify to what extent features interact with each other. This cannot be done for neural networks, as the permutation explainer is not capable of producing these interaction values.

The following Subsections present the results of the deeper analysis into feature use. Only the most informative dependence plots are shown. All dependence plots can, however, be found in Appendix B.

5.8.1. Prediction of Water Content

In order to be able to understand the interpretation of the depth feature, models A2-NN and A2-RF are further analyzed.

Depth

The dependence plots for depth, shown in Figures D.1 and D.4, indicate that some sensor depths were associated with higher SHAP values than others. For example, the most shallow sensor at 1 meter depth was found to generally decrease the prediction of θ, while the sensor at 2 meters depth was not found to be important by the neural network. The neural network and random forest model found different patterns for interpretation of depth, both without finding a strong correlation. For the random forest model, this indicates that the depth feature, being the most important one, is not influencing the output through a generalized feature, but rather through a pattern found within the limited dataset.

Net cone resistance q_t

For the net cone resistance, model A2-NN showed varying interpretations of the raw features, since some correlations were found to be positive, and some negative. Some correlations had a non-linear shape, while other had linear shapes. The strongest correlations were that of $q_{t,\Delta z=-4}$, which had a non-linear, negative correlation, and that of $q_{t,\Delta z=-10}$, which had a linear positive correlation.

Model A2-RF found a weak, negative correlation, which seemingly clustered the feature values into two groups: one below q_t = 1000 kPa, leading to increased predictions of θ, and one above q_t = 1000 kPa, leading to decreased predictions of θ. This pattern was the strongest for q_t at Δz = 0, and became weaker as Δz decreased up to -10. This could indicate that the zone of influence for q_t was closest when Δz was equal to 0 centimeters.

Sleeve friction f_s

For model A2-NN, some strong correlations were observed in the dependence plots, however, it was again the case that some were positive (at Δz equal to 0 and 4), and others negative (at Δz equal to 6 and 8). Since some strong correlations were found, it could indicate that f_s is easier to interpret than q_t.

Model A2-RF, however, did not find correlations as strong as model A2-NN did, but did find a consistent correlation, which was negative and non-linear and got stronger as Δz increased from 0 to 10. The zone of influence of f_s could be the closest to Δz = 10 centimeters.

Interaction between features

Through the tree explainer, the interactions between features can be quantified for the random forest models. To assess the extent of feature interaction, the absolute SHAP interaction values between features are averaged. The resulting heatmap is shown in Figure 5.8. The values on the diagonal represent the importance of associated feature without interaction with other features. A negligible amount of interaction between features is observed.

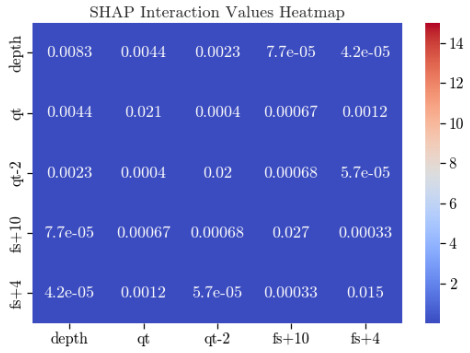


Figure 5.8: Heatmap interaction plot A2-RF

5.8.2. Shear Strength Predicting Models

The models predicting s_u performed well, especially compared to models predicting θ. Hence, the working of the models is especially of interest. The best performing model was the neural network. The dependence plots can be found in Figures D.7 and D.8.

Net cone resistance q_t

The net cone resistance was of little importance to model B7-NN, while it was the second most important features for model B7-RF. The correlations found by the two models both find a positive correlation for q_t, however. In case of the neural network, a near-perfect linear correlation is found, while the random forest finds a non-linear correlation.

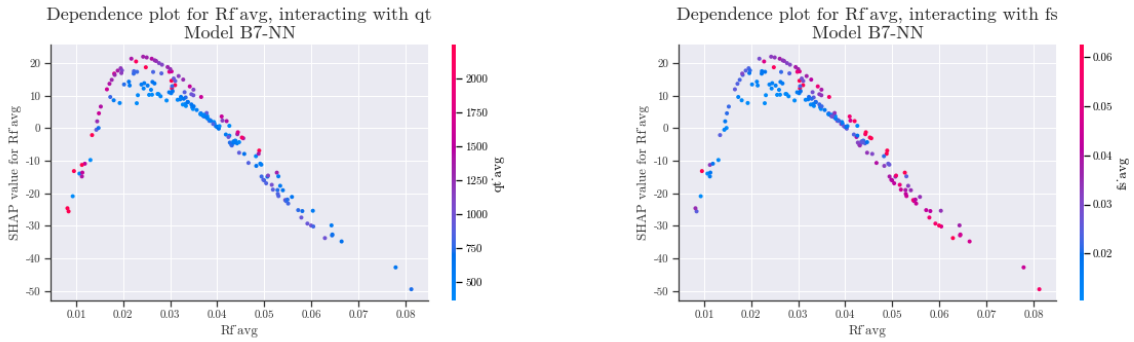
Sleeve friction f_s

Sleeve friction as a feature was important to both the neural network and random forest model. The neural network again found a near-perfect linear, positive correlation, while the random forest found a linear, slightly positive correlation between f_s and s_u.

Friction ratio R_f

The friction ratio was the most important feature for B7-NN and the fourth most important for B7-RF. The dependence plots show that the neural network found a very clear correlation for R_f, which follows a non linear shape. A value of R_f equal to around 2.5% would lead to the largest increase in the predicted value for s_u. No clear correlation was found by the random forest.

As an interesting correlation is observed for the R_f feature, the feature interaction between R_f, q_t and f_s is further investigated. The dependence plot is again given, but this time the colors of the data points represent the feature value of either q_t or f_s. Results are shown in Figure 5.9.



(a) Feature interaction with q_t.

(b) Feature interaction with f_s.

Figure 5.9: Dependence plots Rf for model B7-NN showing feature interaction with qt and fs..

The plots show, that low R_f values are mostly related to high values of q_t, while high R_f values are mostly related to high f_s values. These are the regions where the R_f feature is causing a decrease in the predicted value of s_u. The region where R_f is causing an increase of the predicted value of s_u (i.e., around R_f = 2.5%) consists of low and high values for both q_t and f_s.

Weather features

Monthly average weather features were unimportant for both models. However, model B7-NN found linear correlations for both precipitation and evaporation, which were negative and positive respectively. No clear correlation was found by the random forest models.

The short term weather parameters W was unimportant for the neural network, but was the third most important for the random forest. The neural network again found a near-perfect linear, positive correlation. No clear correlation is seen in the dependence plot for the random forest model.

Interaction between features

For model B7-RF, the SHAP interaction values were once again calculated, which are averaged and shown in the heatmap of Figure 5.10. It should be noted that the SHAP interaction values scale by the SHAP values, which are scaled by the target parameter, hence the values are much different than those in Figure 5.8. To properly visualize the interaction values, the colorbar was set to have a maximum at 15, which is why the intrinsic interaction values of q_t and W are colored the same, although the values are different.

In general, some interaction between features is observed. Especially W interacts a lot with other features, while also being important without the interaction with other features. The most important feature without interactions with other features is q_t. It is also observed that f_s interacts with both q_t and f_s to a larger extent than any of the weather features.

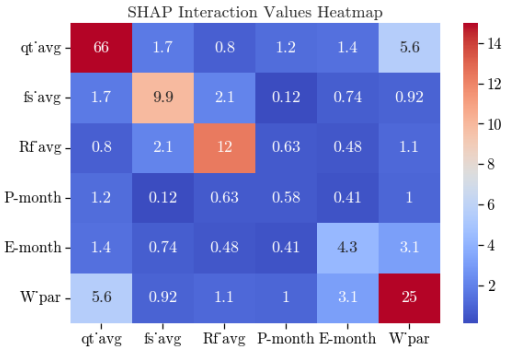


Figure 5.10: Heatmap interaction plot B7-RF

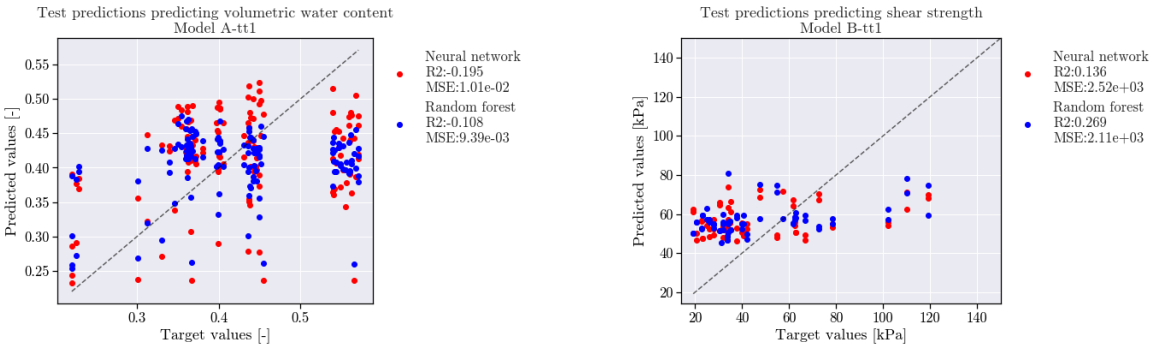
5.9. Transferability Test

In order to assess the generalizing capabilities of the models, a transferability test is performed in two fold. In this test, the input features of the best performing models are used. The training and testing sets are changed, however, and thus hyperparameter tuning is once again performed. The training set consists of all the data from one site (e.g. Maasdijk, Oijen) and the test set of the data from the other site (e.g., IJsseldijk, Westervoort). The test performance is assessed. Then, the training and test sets are swapped, models are trained, and test performance is assessed again.

For models predicting θ the following features were used: q_t, f_s, R_f, P_{month}, E_{month} and W, as was done for models A6-NN and A6-RF, the best performing A-type models. For models predicting s_u, the following features were used: $q_{t,avg}$, $f_{s,avg}$, $R_{f,avg}$, P_{month}, E_{month} and W, as was done for models B7-NN and B7-RF, the best performing B-type models.

In the first fold, training data are from Maasdijk, Oijen and test data from IJsseldijk, Westervoort. The models are referred to as model A-tt1 and B-tt1. The hyperparameters are given in Table B.8.

Test results are given in Figure 5.11. No accurate predictions are made by any of the models, and performance decreased with respect to the best performing models.



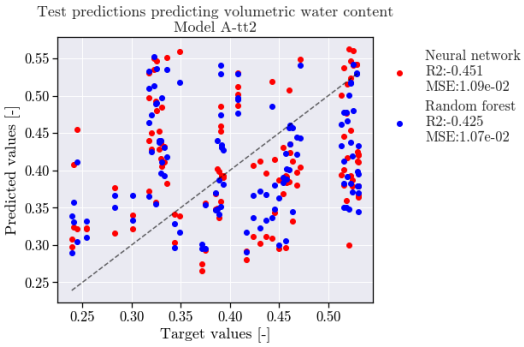
(a) Model A-tt1.

(b) Model B-tt1.

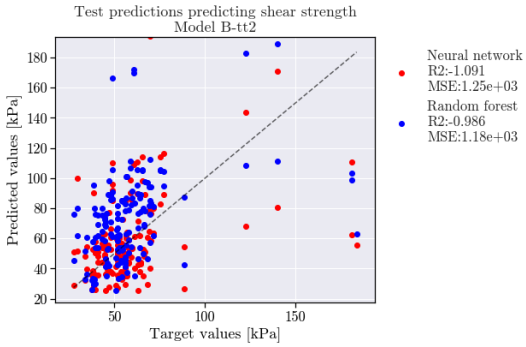
Figure 5.11: Test results models A-tt1 and B-tt1.

For the second fold, the training data consisted of data from IJsseldijk, Westervoort and test data of data from Maasdijk, Oijen. The hyperparameters are given in Table B.9.

The test results are shown in Figure 5.12. Performance decreased even further compared to the first fold of the transferability test, as no accurate predictions are made.



(a) Model A-tt2.



(b) Model B-tt2.

Figure 5.12: Test results models A-tt2 and B-tt2.

6

Discussion

The results presented in Chapter 5 included intermediate observations and conclusions that were necessary in order to progressively change model features. The main observations are summarized and discussed below.

Model performance

Test performance was found to be inconsistent. Sometimes, neural networks performed better, while at other times random forests performed better. However, one should consider that neural networks were validated using K-fold validation, while random forests were not. Though it is difficult to compare methodologies directly, it could be considered that random forests were able to use the limited dataset more efficiently, as they produced a more consistent test performance. Neural networks might have a higher potential performance, but need knowledge-based feature engineering to improve. Another possibility to improve model performance would be to increase the dataset size.

Neural network feature importance

Overall, f_s, R_f, and depth were the only important features when predicting either θ or s_u according to the neural network models. In fact, some neural networks made predictions solely based on f_s. The SHAP analysis showed that neural networks made little use of q_t as a feature, which is an unexpected result. Commonly, CPT interpretation never bypasses q_t, but rather f_s is the bypassed parameter.

From a physical point of view, two main reasons could be thought of as to why q_t is bypassed. The first reason is, that q_t does not possess useful information. The second reason is, that f_s is able to pass the same physical information to the model, and more. Since the first reason cannot be the case, since q_t is known to relate to both θ and s_u in literature and the dataset, the second reason must be the case. The neural network models thus make an interesting observation that f_s is more informative for them in the use cases of this thesis than q_t.

Weather features were also not used by the neural networks. Similarly to q_t, it is known that weather influences do relate to θ and s_u, thus it must again be the case that f_s already possesses the important information stored in weather data. This is an important observation, as it means that on site monitoring of weather effects is not a necessity when CPT data is available.

Although this physical interpretation would lead to new insights, it should be considered that modeling limitations could actually be the true cause for the neural networks' remarkable feature interpretation. It should, namely, be considered that low model performance is achieved when f_s is the only feature on which the model bases its predictions. It could mean that feature interactions inside the neural network models are scarce, leading to overfitting on one feature, while performance could be improved if more features were used. This could be solved by adding more training data, or increasing the regularization term. It could also be an problem inherent to the neural network modeling method, in which case, no other solution exists but to change to a different modeling method.

Random forest feature importance

Random forest models consistently found meaning in q_t and f_s the prediction of both θ and s_u. However, prediction of θ became difficult after the removal of depth as a feature, while prediction of s_u stayed relatively consistent throughout the testing of different input layers.

In the prediction of θ, f_s was generally more important than q_t. This suggests that fluctuations of θ are represented to a greater extent in f_s than in q_t. An interesting observation is found in the SHAP feature importance summary plots for random forest models predicting θ. Consistently throughout all the models, it can be noticed that q_t has a large, negative influence on the predictions of θ when the feature values are high, while f_s has a large, positive influence on the predictions of θ when the feature values are low. To put this in other words: especially high q_t values and especially low f_s values are important when predicting θ. The former can also be observed in the data analysis in Chapter 4, where high q_c values are found at shallow depths, when θ values have dropped due to seasonal influence. The random forest makes less use of the fact that, at these shallow depths, higher f_s values can also be found when θ drops. Instead, it notices how low f_s values should correlate to high θ values, as these are found at deeper depths.

It is in line with expectations that q_t was the most important feature for s_u prediction, but it is a new finding that f_s can additionally be used by the model. Correlations of f_s are mostly similar to those of q_t, but were apparently still of additional value to the model. It is a meaningful observation that the test performance stayed relatively consistent, as this demonstrates the power of these two core features. By testing the model using different additional features, it was found that R_f provides additional information to the model, so that performance increases. This could be caused by the feature interaction becoming easier, as is shown in Figure 5.10.

Depth as a feature

Depth was removed as a potential feature, leading to difficulty in the prediction of θ. One could wonder whether this was justified, as depth was not always an important feature when it was used as a feature in Section 5.1. Possibly, in those models, the other features overshadowed the bias introduced by the depth feature. However, since the test performance increased so drastically in models A3-NN, A3-RF, and B3-NN when depth was the only important feature, the legitimacy of models using depth should then always be additionally tested.

It is logical that solely depth can predict θ very well, which becomes apparent when Figure 4.3 is observed. The 8 sensors are placed at only 7 unique depths, meaning that each depth can most of the time directly be linked to the sensor. Some sensors measure very little water content fluctuations. Hence, depth can directly inform the model about a good, educated guess. Since the depths are less discrete when predicting s_u, as seen in Figure 4.11, this risk does not occur in the same extent for the prediction of s_u.

Although depth itself was removed as an explicit feature, it should be noted that depth still implicitly was included in features. q_t and W both include depth in their definition, and capture the most important effects depth is expected to have on the output.

Taking into account the above considerations, and the fact that the performance of models predicting s_u was able to be improved even without depth, it is thought that the removal of the depth feature was justified. However, some valuable information is lost, as depth can give machine learning models information as to where the initially unsaturated layer is. A better and more fair parameter could be the depth, with respect to the groundwater table. Using this parameter would, however, mean that a constant measurement or accurate prediction of the groundwater table is required, which both were no options in this thesis.

Another consideration that should be made, is regarding the limitations of SHAP. During the research, it was assumed that the results of the SHAP analysis are always true. However, it could be considered illogical that model B4-NN performed much worse than B2-NN, since, according to the SHAP analysis, they mostly use the same important features, except for the small contribution of depth to B2-NN. Possibly, SHAP is not analyzing the interpretation correctly, and the contribution of depth to model B2-NN might be larger than it is suggesting.

Feature engineering

The engineering of weather features lead to a performance increase for 3 out of the 4 models, being especially useful to the prediction of θ. The engineering of CPT features was only beneficial to models predicting s_u, but did cause the greatest performance increase observed going from model B6-NN to B7-NN, which ultimately was the best performing model.

The performance increase can be explained by two contributing factors. The first, is that the size of the input layer decreases, which reduces data sparsity associated with the curse of dimensionality. The second is, that the features are able to retain the most important information going from raw to engineered features. Based on the results, it is not possible to distinguish how much of the performance increase is caused by the factors separately, however, both are considered to be the case.

The fact that the engineering of weather features was more beneficial to the prediction of θ than the prediction of s_u could indicate that weather-related information in general is more important to models predicting θ, even though weather features were relatively unimportant compared to CPT features for these models. Possibly, the engineered weather features were able to help the models interpret the CPT features more easily, leading to an indirect importance of the weather features. Though, it could also be the case that the large size reduction of the input layer was the main reason the models increased in performance. However, since no importance was found in raw weather features, no conclusion can be made regarding the importance of time lag associated with weather data and subsurface consequence.

Feature engineering of CPT features did, however, lead to a performance decrease for the prediction of θ. This indicates that valuable information was lost by averaging the CPT parameters over the chosen depth interval. Therefore, it could be concluded that local interactions along the depth interval that was considered are actually important in this problem. In other words, the results suggest that the zone of influence related to CPT parameters that directly relates to the volumetric water content is at smaller than the chosen depth interval of 10 centimeters.

For models predicting s_u, it could be reasoned that CPT features already capture the most important information stored in weather features, since no direct performance increase was observed after feature engineering of weather features. Still, the engineering of weather features seemed to be useful for interpretation of weather data, and for finding interactions between features, as the best performing random forest model (B7-RF) found importance in those weather features and found a lot of interaction between W and other features.

As engineering of CPT features by averaging them over a 10 centimeter interval lead to the best performing s_u-predicting models, it can be considered that local interactions are less important when predicting s_u, and the simplification helps the models understand the core information of CPT features. This makes sense, as s_u is measured using field vane tests, which themselves have a height of 8-20 centimeters, so the zone of influence is at least 8 centimeters along the depth interval. Averaging over this interval seems to be a good way of dealing with CPT parameters.

Raw CPT parameters and zone of influence

The inconsistency of interpretation of q_t and f_s when raw CPT features were used can be explained through two reasons. Reason one, which relates to the modeling, is that the most important measurements are used primarily to establish a baseline prediction. Then, the less important features are used to regularize this prediction. The second reason relates to a physical meaning. It could. namely, be the case that the negative correlations found at measurements where Δz = 0, +2, and +4 was caused because of local interactions between those measurements and measurements where Δz = +6, +8, and +10. It could be the case, that suction mechanisms are found as a pattern, where high suction surrounding a point causes a decrease of θ at that point. No conclusion can be made based on the results if these two reasons are truly happening and to what extent.

Throughout all of the testing, it was found that, for q_t, the overall most important measurement originated from Δz = 0 cm. For f_s the overall most important measurement originated from Δz = +10 cm, where Δz refers to Figure 4.15. This result means that it would always be important to correct CPT measurements so that the zones of influence of q_t and f_s align, which is not commonly done.

With regard to CPT cone dimensions, it makes sense that this Δz values is around 10 centimeters for f_s. However, as this was the maximum Δz value investigated, it could be the case that even larger values would lead to a more appropriate correction of CPT parameters. However, since f_s measured at Δz = 8 cm was often second most important, or even more important in the case of model B5-NN (Figure C.5c), there is no reason to believe that much larger values of Δz would lead to a better correction.

R_f as a feature

Demonstrated by the performance increase of models A6-NN and A6-RF, it can be concluded that R_f is a valuable parameter to machine learning models predicting θ. This aligns with expectations based on the data analysis, as shown in Figure 4.10. It could be considered that R_f helps understand the complexity and non-linearity at hand, as it does not actually provide new information, since it is merely a ratio between f_s and q_t, which was previously already provided to the previous models.

The addition of R_f initially lead to a performance decrease of model B6-NN. It could be the case, that the expansion of the input layer was the cause for this, due to the curse of dimensionality. After feature engineering and reduction of the size of the input layer, model B7-NN became the best performing model, using mostly f_s and R_f together. This model found an interesting correlation for R_f in the dependence plots shown in Figure 5.9, which could be a useful new insight how to make use of R_f in the prediction of s_u. Although it should be considered that SHAP value and target value are not necessarily correlated. SHAP value merely explains how the model's predictions are influenced by a feature, with respect to its baseline prediction. However, since the correlation is very strong for R_f in Figure 5.9, it can be concluded that the model finds a meaningful way to interpret this parameter.

Feature interpretation

Although models A6-NN and A6-RF did not perform well, their feature interpretation in Section 5.8 shows interesting results. For model A6-NN, a strong, positive linear correlation was found for R_f and W between the feature and SHAP value, which are aligned with expectations. However, a positive correlation was also found for f_s by model A6-NN, which contradicts expectations. This could be a reason as to why the model's predictions are bad. Model A6-RF performed somewhat better, and indeed did find a negative correlation between f_s's SHAP and feature value. This correlation followed a non-linear shape, which could even be compared to the shape of a soil water retention curve (Figure 2.4).

Next to that, weather features were interpreted more clearly by the neural network than the random forest model. It was observed that correlation between SHAP and feature value for precipitation and evaporation features followed a bow-tie shape, with the center of the bow-tie crossing the x-axis at P_{month} = 2.2 mm and E_{month} = 1.5 mm. These feature values thus describe the point at which precipitation/evaporation causes an increase/decrease in prediction of θ. The values differ by 0.7 mm, which is an interesting observation, as one could think that, if precipitation is higher than evaporation, an increase in θ should be expected. However, this difference of 0.7 mm could be attributed to precipitation that does not end up in the investigated soil body. This water is for example, runoff from the dike, transported to the subsurface below the dike, or absorbed by plants.

All in all, it can be concluded that neural network A6-NN was better at interpreting weather features, while random forest A6-RF was better at interpreting CPT features. As both models found CPT features the most important, this resulted in better performance of random forest models predicting θ. Figure 5.8 indicated little interaction between features, which could be a reason why the models performed badly. Possibly, it could be solved by adding an additional feature, which is able to help understand the interaction between features. This interactive feature could be useful if depth-dependent, though this was not a possibility using the provided dataset, as discussed earlier.

The functioning of the models predicting s_u differed analyzing the dependence plots and interaction heatmap in Section 5.8.2. Model B7-NN was able to interpret strong correlations for each feature individually, while these correlations were weaker by model B7-RF. The performance of B7-RF was close to that of B7-NN, however, which could be attributed to the fact that features were able to interact with each other.

Model B7-NN found strong, linear correlations between SHAP and feature value for q_t, f_s, P_{month}, E_{month}, and W, and a strong non-linear correlation for R_f. Even though the neural network found

these correlations, it only used f_s and R_f for predictions. This strengthens the idea that the unimportant features (q_t and weather features) were bypassed because they did not provide new information with respect to f_s and R_f. Since all but R_f's correlations were linear, this makes sense, as this linearity can be summarized in one feature, which was, in the model's case, f_s. The non-linearity, which is expected to be the case for the problem, was introduced by the R_f feature, thus this feature did provide new information and model performance increased.

The SHAP interaction values shown in Figure 5.10 shows that the random forest model was able to find meaningful interactions between some of the features. Especially q_t and W were important features, not only when isolated from other features, but also when interacting with them. Apart from W, weather features show little interaction with CPT features, which could be a reason why they are less important for the model's predictions.

Transferability

The transferability tests showed the weak generalizing strength of the models when trained on data from one site and tested on another. Two main reasons could be thought of why this is the case.

Reason one, is that the test sites could differ to a large extent in soil characterization, weather effects, and/or measurements. However, as discussed and shown in Section 4, this is not the case. The second reason, is that the location along the dike (crest vs. berm) is very important with regard to the inputs and outputs used in the models. This second reason makes sense, as the different locations along the dike intersection have a large difference in height with respect to the water table, causing the initially unsaturated layer to be much larger on the dike crest than on the inner berm. Hence, an additional feature is needed, that describes this discrepancy, in order to generalize a model that trains on data from one site and is used to predict for another site.

Another solution on how to build more generalized models and improve on the transferability test, is to combine data from different sites, and increase the size of the training dataset. This is what was also done in models 1-7, which combined data from Oijen and Westervoort in one dataset. Since the performance of these models were much better than those used in the transferability test, it could be extrapolated that adding even more data from different sites would increase model performance even further. This could be a next research step, if machine learning models are to be applied in actual shear strength determination where accurate predictions are key in order to properly assess macro stability for many different dikes. Additionally, it should be noted that, in this thesis, all data originated from two primary dikes, thus transferring these models to organic dikes could prove difficult, as correlations between CPT features and θ or s_u are different for these types of soils.

7

Conclusion

In the commonly used methodology of shear strength determination, the cone resistance is the only CPT parameter used in a linear relation. As this produces inaccurate predictions, this thesis focused on the exploitation of CPTs to improve the determination of shear strength throughout seasonal effects, with special interest in the use of sleeve friction, since this is an underutilized parameter, even though it is always measured in parallel with cone resistance.

The following main research question was posed.

How can machine learning models be used in the determination of shear strength of unsaturated layers in a dike using cone penetration test and weather data?

Through knowledge-based feature engineering, machine learning models were able to improve the shear strength determination compared to conventional methods. Sleeve friction significantly contributes to the determination of s_u, as demonstrated by its importance in both neural network models and random forests. In fact, neural network models were able to make their predictions without using q_t, and often solely based their predictions based on f_s. The best performing model, however, combined f_s with R_f to make accurate predictions of s_u. This thesis demonstrated the value of feature engineering techniques, assessed using SHAP, while exploiting CPT and weather data for improved shear strength prediction.

These conclusions were made with the help of five sub-questions. The questions and answers were as follows.

How are CPTs typically interpreted and what relations to seasonal influence have been found?

Typically, CPTs were found to be widely used for various geotechnical purposes, especially for strength determination problems. Commonly, the cone resistance was used for this, while sleeve friction was often considered less accurate. However, different sources showed a relation between CPT parameters and water content of the soil sample. It was found that the higher the water content, the lower the cone resistance and sleeve friction, being the main reason why seasonal effects influence CPT measurements. Sleeve friction was also found to correlate to remolded shear strength. Different interpretations of adhesion with respect to a steel-soil shear interface with relation to wet and dry conditions were observed.

To what extent is seasonal influence present in the dataset, consisting of CPT, FVT and weather data?

The data analysis showed that seasonal influence was found in CPT and FVT measurements at both test sites, though to a different extent. In CPTs, this seasonal influence was mostly found in the top 1.5-2 meters, defining the initially unsaturated layer. The data analysis also showed that the current methodology for shear strength determination produces poor results at the Oijen test site, where data is measured at the dike crest and seasonal influences are very relevant. At the Westervoort test site, seasonal influence is less relevant, and the current methodology produces better results.

Can machine learning models be used to make predictions regarding the hydro-mechanical state of the dike?

Machine learning models were able to improve the prediction of s_u using CPT and weather data. The best performing models produced decent predictions and used engineered features of q_t, f_s, R_f, precipitation, and evaporation. Machine learning models were unable to make accurate predictions of θ using CPT and weather data. Predictions were accurate when depth was used as an input feature, but this was a deceiving result caused by the limitations of the dataset. It is necessary to train a model using data from different test sites in order to produce generalized results, as model performance was poor when training and test data originated from two different sites.

How do machine learning models interpret CPT and weather related features?

It was found that the most influential f_s measurement, indicating the location of the zone of influence, followed behind by 10 centimeters compared to the CPT tip. Friction ratio can be used to improve model interpretation of f_s showcased by the improved neural network predictions. A non-linear interpretation of R_f was found by the neural network model. Weather features were often unimportant to the models, possibly indicating that CPT parameters already capture the essence of seasonality, and the addition of weather features does not provide new information.

To what extent can machine learning models be improved through knowledge-based feature engineering?

Feature engineering, paired with dimensionality reduction, led to an increase in test performance when predicting s_u, especially for neural network models. Random forest models are more capable of dealing with high dimensionality of the input data. Weather data became more important through feature engineering, after which they were able to interact with other features to improve test performance of random forest models. Overall, it was concluded that the input features of these models should always be picked carefully, and if possible, reduced to as little features as possible, since this improved the test performance.

7.1. Recommendations

It was noted that the results showed detrimental loss of performance in the prediction of θ when depth was removed as a feature. Alternatives to using depth should thus be explored, which are less likely to introduce a bias while still capturing the same information. It is therefore recommended to use depth with respect to groundwater table rather than depth with respect to surface level. As the groundwater table is continuously changing, this parameter would remove the discretization problems that arose with the depth feature. However, it is costly to know or accurately predict the ground water table. Possibly, a prediction model could be made based on river water level and water levels on the inside of the dike. A more expensive, but likely more accurate method would be to monitor using standpipes.

Poor generalization was observed in the transferability test, which was to be expected following poor test results for some of the models that were used. However, it could have also been the case that that the discrepancy between cross-sectional dike locations between the two test sites was the cause. One possible fix for this also was to introduce a generalizing feature, such as the depth with respect to the groundwater table. The discrepancy would in this way possibly be unnoticed by models. Otherwise, it is also recommended to increase the size of the training dataset, while ensuring to include data from different test sites in order to achieve generalization.

Although clear correlations were found by model B7-NN between feature and SHAP value, it is important to note that that does not necessarily mean that there is also good correlation between feature and target value. It does mean, however, that the important features, f_s and R_f, can be considered valuable and interpretable by the model. Therefore, it is recommended to further research what the correlation is between these parameters and target parameters, such as θ and s_u. This could be done through laboratory tests. Ideally, since interactions with cone resistance are additionally of interest, these correlations would be investigated in calibration chamber testing, reproducing CPT results. Soil samples with different water content would be used to understand the direct correlation between θ and CPT parameters.

In this thesis, machine learning models did not consist of any physical knowledge; training was purely data-driven. It has, however, been demonstrated that physical information can be of great value for machine learning methods, in order to improve performance and interpret results. In this thesis, this was adequately done through feature engineering and importance assessment. However, it would be interesting to see if performance can even be improved further if models become physics informed. If the physical relation between CPT parameters and hydro-mechanical parameters becomes more clear, it could be considered to use this relation in a physics-informed neural network.

References

Almasoudi, R., Abuel-Naga, H., & Daghistani, F. (2023). Effects of dry density and moisture content on the kaolin–brass interfacial shear adhesion. *Applied Sciences*, *13*(20), 11191.

Arcadis. (2020). *Consequentieanalyse initieel niet verzadigde zone* (tech. rep.).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Campanella, R., Gillespe, D., & Robertson, P. (1982). Pore pressures during cone penetration testing. *Proceedings, 2nd European Symposium on Penetration Testing*.

Ceccato, F., & Simonini, P. (2017). Numerical study of partially drained penetration and pore pressure dissipation in piezocone test. *Acta Geotechnica*, *12*, 195–209.

Chaya. (2020). Random forest regression. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

DeJong, J., Cargill, P., & Frost, J. (2001). Effect of surface texturing on cpt friction sleeve measurements. *Journal of Geotechnical and Geoenvironmental Engineering*, *127*, 105–201.

Elsawy, M. B., Alsharekh, M. F., & Shaban, M. (2022). Modeling undrained shear strength of sensitive alluvial soft clay using machine learning approach. *Applied Sciences*, *12*(19), 10177.

Farrar, J., Torres, R., & Crutchfiedl, L. (2008). *Cone penetrometer testing, scoggins dam, tualatin project, oregon.* (tech. rep.). Engineering Geology Group Bureau of Reclamation, Technical Services Center.

Giacheti, H., Bezerra, R., Rocha, B., & Rodrigues, R. (2019). Seasonal influence on cone penetration test: An unsaturated soil site example. *Journal of ROck Mechanics and Geotechnical Engineering*.

J.J.M. Powell, R. Q. (1988). The interpretation of cone penetration tests in clays, with particular reference to rate effects. *Proc 1st International Symposium on Penetration Testing*.

Keaveny, J., & Mitchell, J. (1986). Strength of fine-grained soils using the piezocone. *Use of In-Situ Tests in Geotechnical Engineering*, *GSP 6*.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. *Advances in neural information processing systems*, *30*.

Koninklijk Nederlands Meteorologisch Instituut (KNMI). (2025). Waarnemingen. https://www.knmi.nl/nederland-nu/weer/waarnemingen

Konrad, J. (1987). Piezo-friction-cone penetrometer testing in soft clays. *Canadian Geotechnical Journal*, *24*, 645–652.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Lunne, T., & Andersen, K. (2007). Soft clay shear strength parameters for deepwater geotechnical design. *Proceedings of the 6th International Offshore Site Investigation and Geotechnics Conference*.

Lunne, T., Robertson, P. K., & Powell, J. J. M. (1997). *Cone penetration testing in geotechnical practice*. Blackie Academic & Professional.

Lunne, T., Robertson, P., & Powell, J. (1997). *Cone penetration testing in geotechnical practice*. Blackie Academic.

Miller, G., Tan, N., Collins, R., & Muraleetharan, K. (2018). Cone penetration testing in unsaturated soils. *Transportation Geotechnics*, *17*, 85–99.

Phoon, K. K., & Kulhawy, F. H. (1999). Characterization of geotechnical variability. *Canadian Geotechnical Journal*, *36*(4), 612–624.

Robertson, P. (1990). Soil classification using the cone penetration test. *Canadian Geotechnical Journal*, *27*, 151–158.

Robertson, P. (2009). Interpretation of cone penetration tests - a unified approach. *Canadian Geotechnical Journal*, *46*, 1337–1355.

Shao, W., Yue, W., Zhang, Y., Zhou, T., Zhang, Y., Dang, Y., Wang, H., Feng, X., & Chao, Z. (2023). The application of machine learning techniques in geotechnical engineering: A review and comparison. *Mathematics*, *11*(18), 3976.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.

Teh, C., & Houlsby, G. (1991). An analytical study of the cone penetration test in clay. *Géotechnique*, *41*, 17–34.

van Duinen, A. (2021). *Shear strength of initially unsaturated soil* (tech. rep.). Deltares.

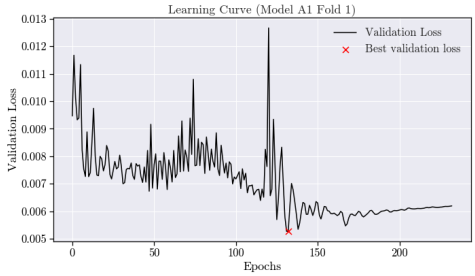
Zein, A. (2017). Estimation of undrained shear strength of fine grainde soils from cone penetration resistance. *International Journal of Geo-Engineering*.

Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., & Ding, X. (2021). Application of deep learning algorithms in geotechnical engineering: A short critical review. *Artificial Intelligence Review*, 1–41.

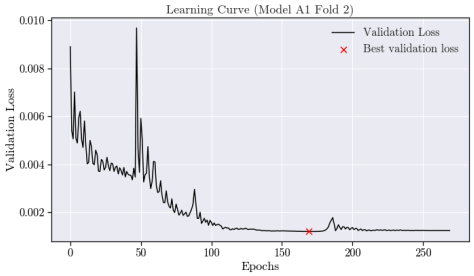
A

Appendix A: Learning Curves

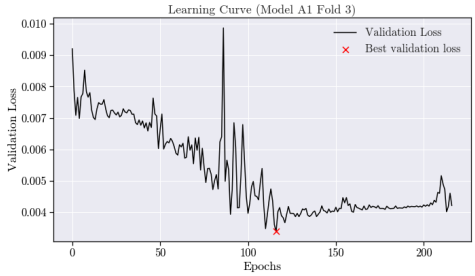
Model A1



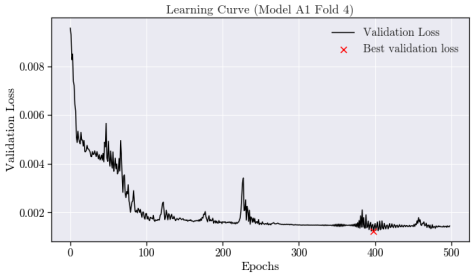
(a) Model A1 Fold 1



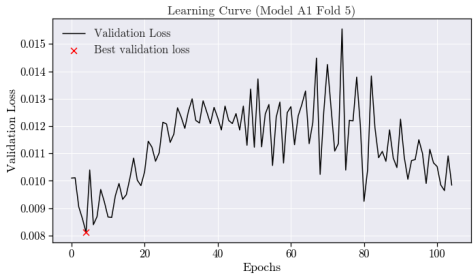
(b) Model A1 Fold 2



(c) Model A1 Fold 3



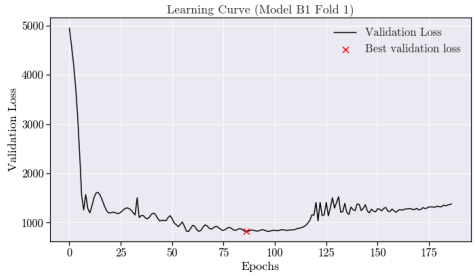
(d) Model A1 Fold 4



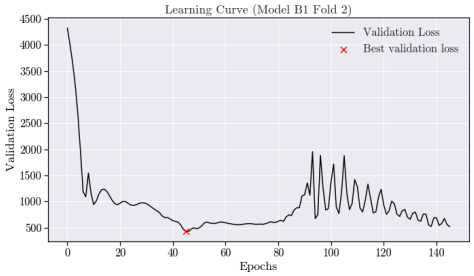
(e) Model A1 Fold 5

Figure A.1: Learning curves for Model A1 across 5 folds.

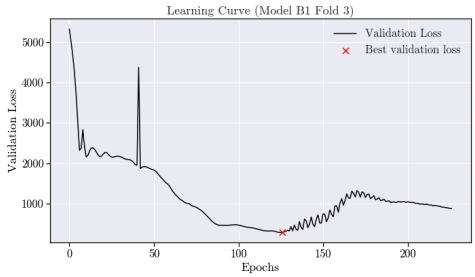
Model B1



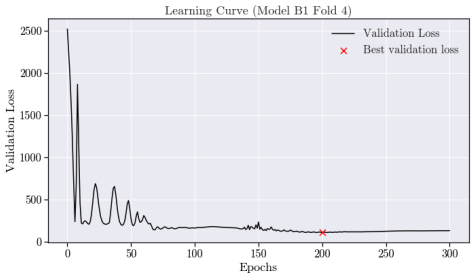
(a) Model B1 Fold 1



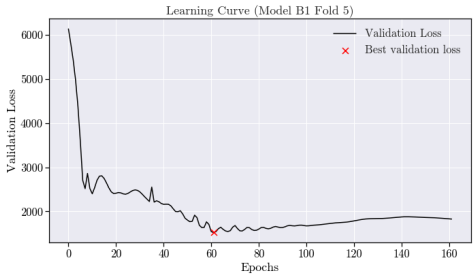
(b) Model B1 Fold 2



(c) Model B1 Fold 3



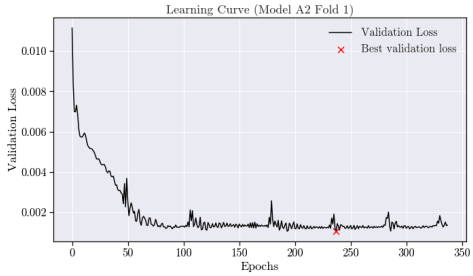
(d) Model B1 Fold 4



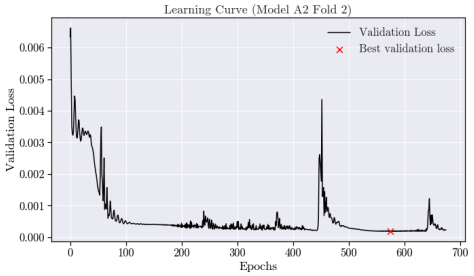
(e) Model B1 Fold 5

Figure A.2: Learning curves for Model B1 across 5 folds.

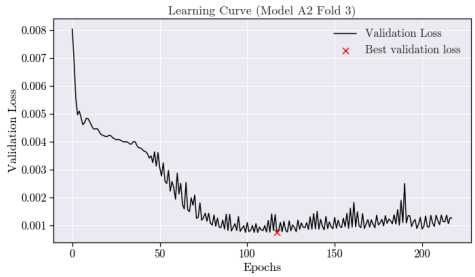
Model A2



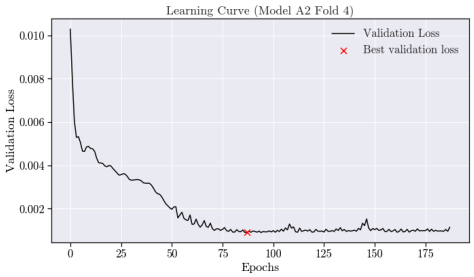
(a) Model A2 Fold 1



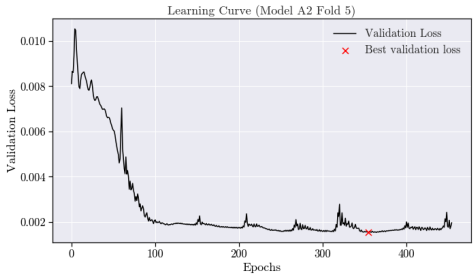
(b) Model A2 Fold 2



(c) Model A2 Fold 3



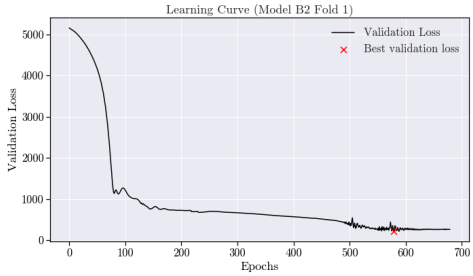
(d) Model A2 Fold 4



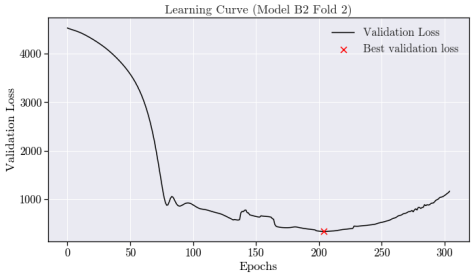
(e) Model A2 Fold 5

Figure A.3: Learning curves for Model A2 across 5 folds.

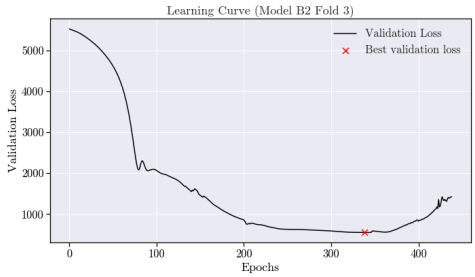
Model B2



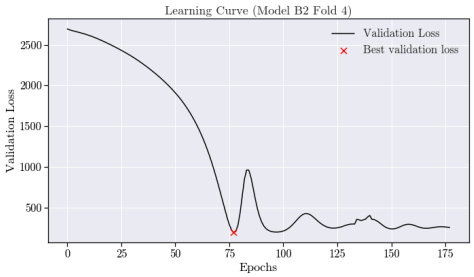
(a) Model B2 Fold 1



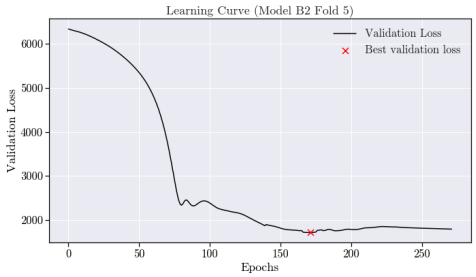
(b) Model B2 Fold 2



(c) Model B2 Fold 3



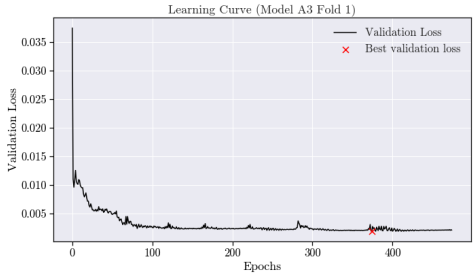
(d) Model B2 Fold 4



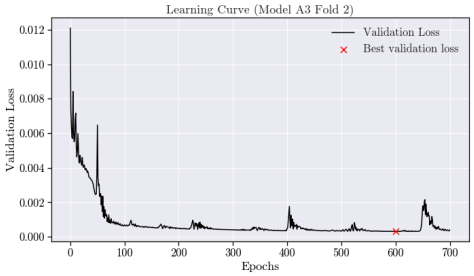
(e) Model B2 Fold 5

Figure A.4: Learning curves for Model B2 across 5 folds.

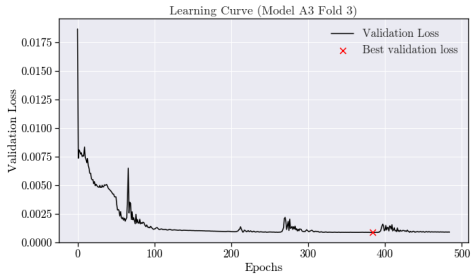
Model A3



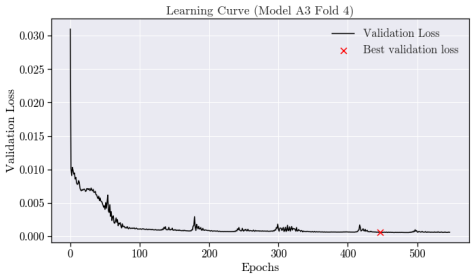
(a) Model A3 Fold 1



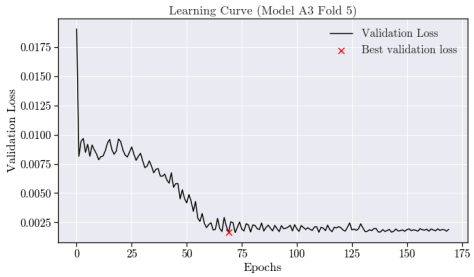
(b) Model A3 Fold 2



(c) Model A3 Fold 3



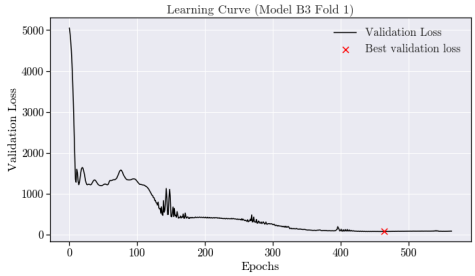
(d) Model A3 Fold 4



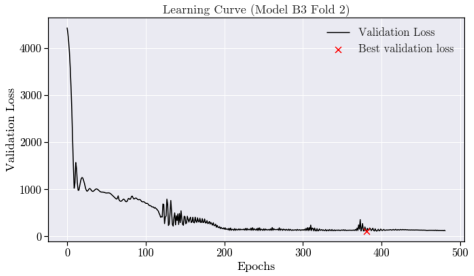
(e) Model A3 Fold 5

Figure A.5: Learning curves for Model A3 across 5 folds.

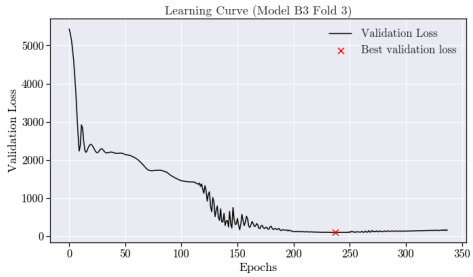
Model B3



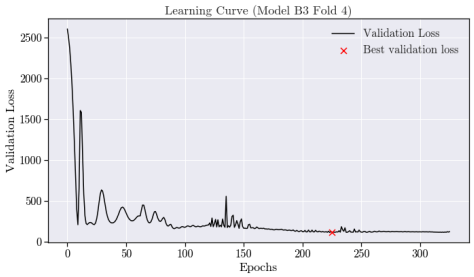
(a) Model B3 Fold 1



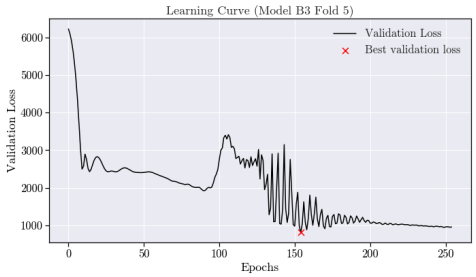
(b) Model B3 Fold 2



(c) Model B3 Fold 3



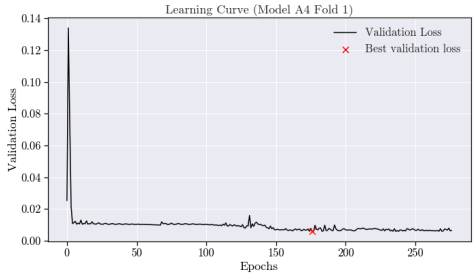
(d) Model B3 Fold 4



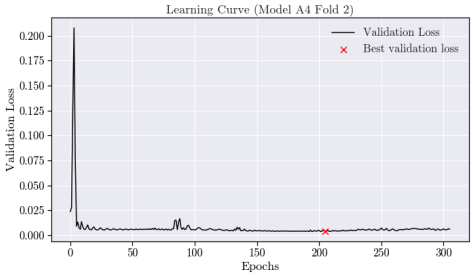
(e) Model B3 Fold 5

Figure A.6: Learning curves for Model B3 across 5 folds.

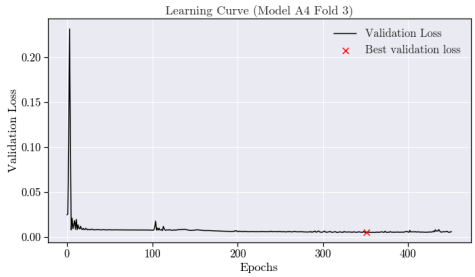
Model A4



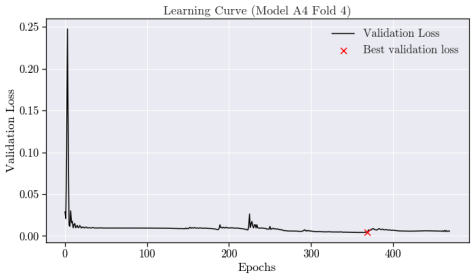
(a) Model A4 Fold 1



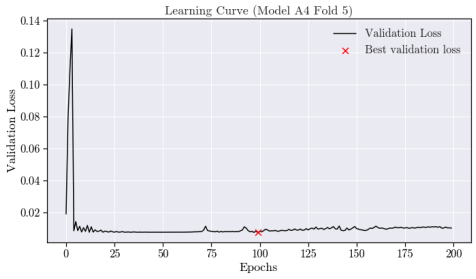
(b) Model A4 Fold 2



(c) Model A4 Fold 3



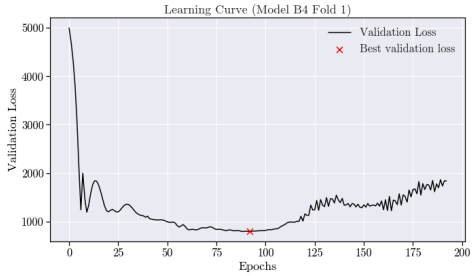
(d) Model A4 Fold 4



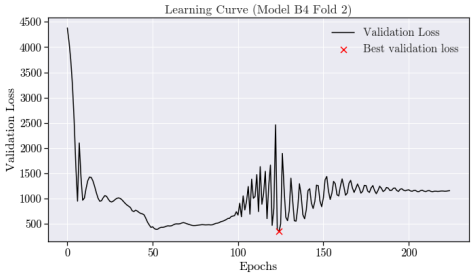
(e) Model A4 Fold 5

Figure A.7: Learning curves for Model A4 across 5 folds.

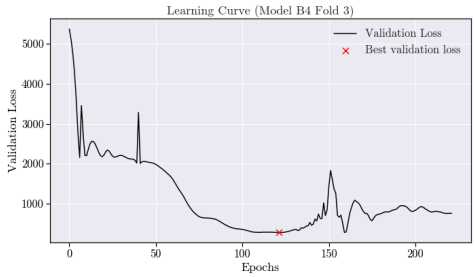
Model B4



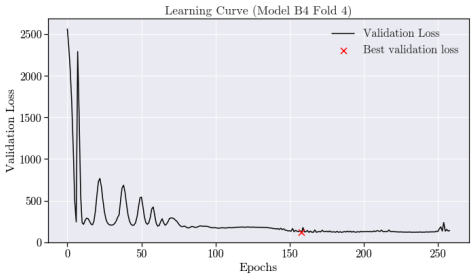
(a) Model B4 Fold 1



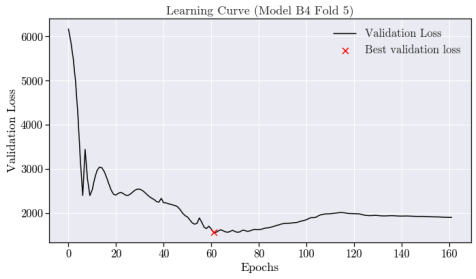
(b) Model B4 Fold 2



(c) Model B4 Fold 3



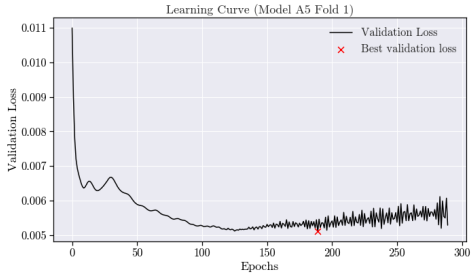
(d) Model B4 Fold 4



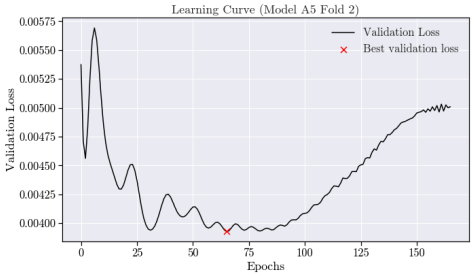
(e) Model B4 Fold 5

Figure A.8: Learning curves for Model B4 across 5 folds.

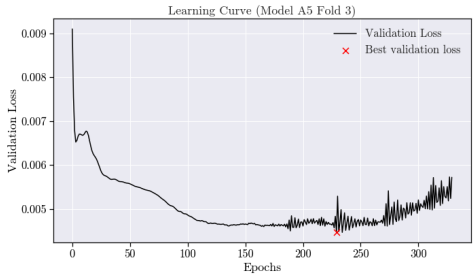
Model A5



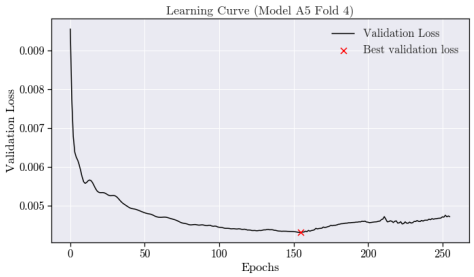
(a) Model A5 Fold 1



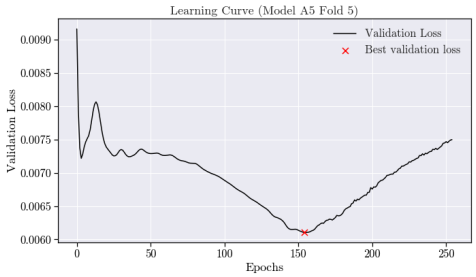
(b) Model A5 Fold 2



(c) Model A5 Fold 3



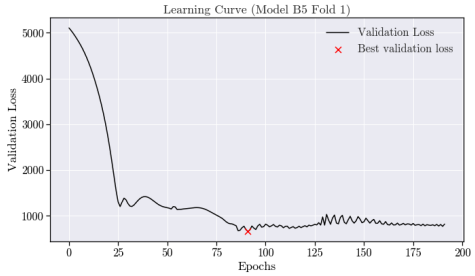
(d) Model A5 Fold 4



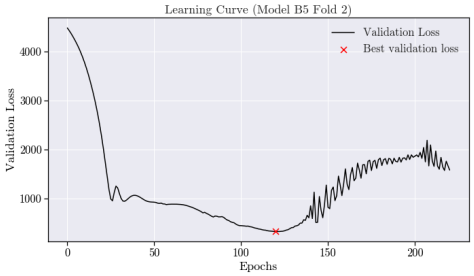
(e) Model A5 Fold 5

Figure A.9: Learning curves for Model A5 across 5 folds.

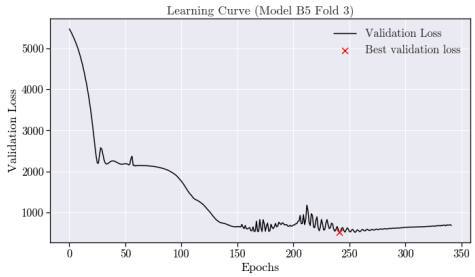
Model B5



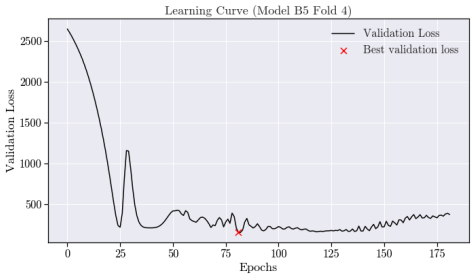
(a) Model B5 Fold 1



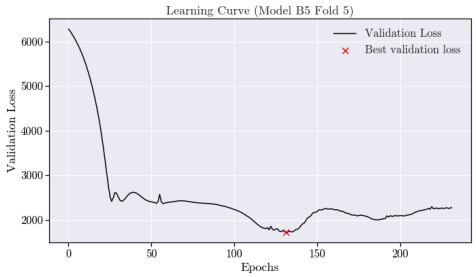
(b) Model B5 Fold 2



(c) Model B5 Fold 3



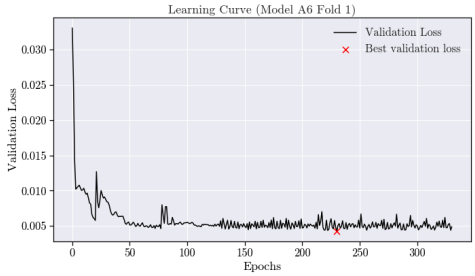
(d) Model B5 Fold 4



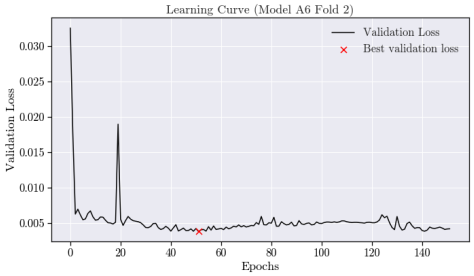
(e) Model B5 Fold 5

Figure A.10: Learning curves for Model B5 across 5 folds.

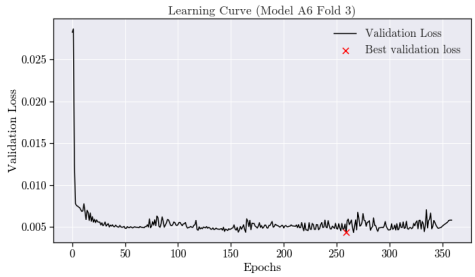
Model A6



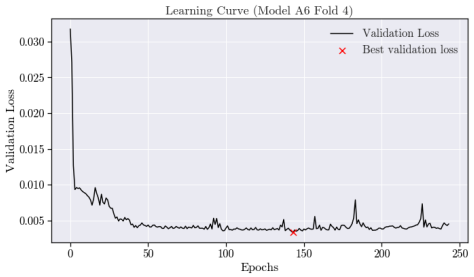
(a) Model A6 Fold 1



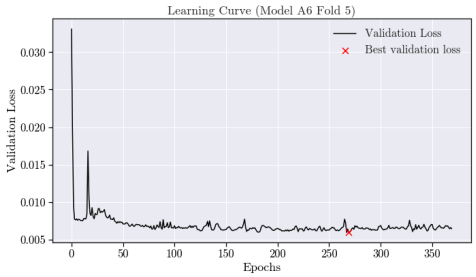
(b) Model A6 Fold 2



(c) Model A6 Fold 3



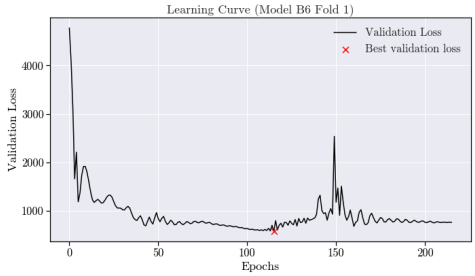
(d) Model A6 Fold 4



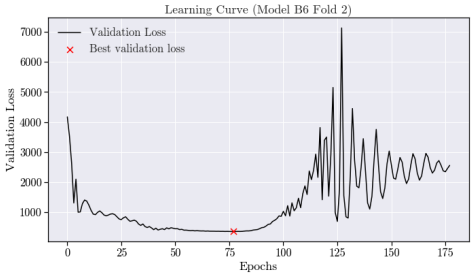
(e) Model A6 Fold 5

Figure A.11: Learning curves for Model A6 across 5 folds.

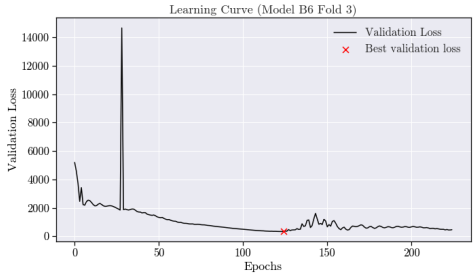
Model B6



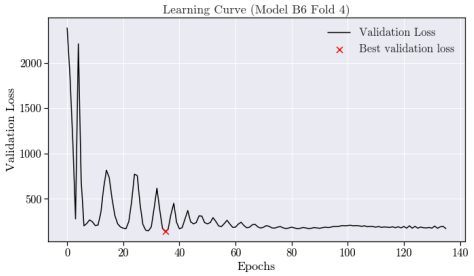
(a) Model B6 Fold 1



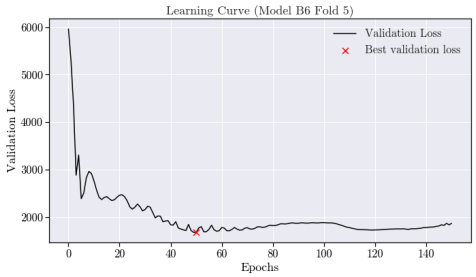
(b) Model B6 Fold 2



(c) Model B6 Fold 3



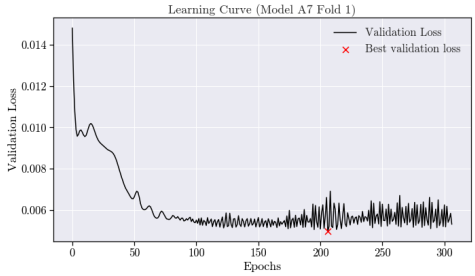
(d) Model B6 Fold 4



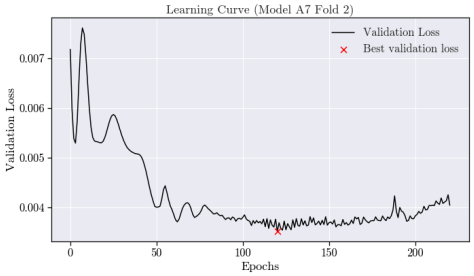
(e) Model B6 Fold 5

Figure A.12: Learning curves for Model B6 across 5 folds.

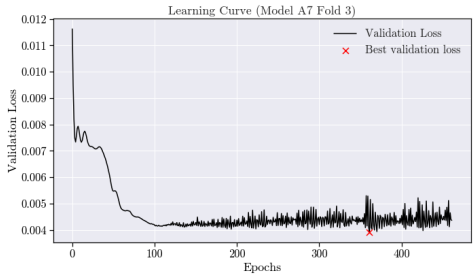
Model A7



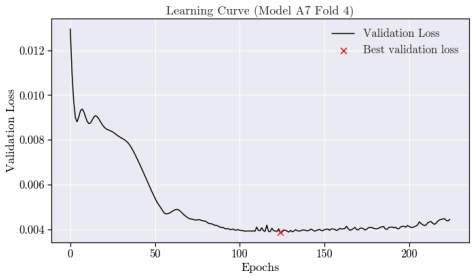
(a) Model A7 Fold 1



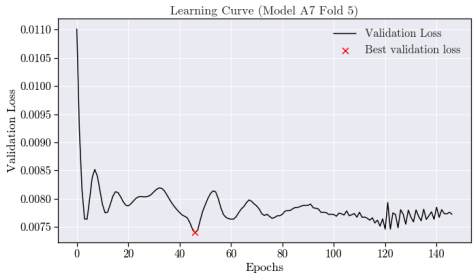
(b) Model A7 Fold 2



(c) Model A7 Fold 3



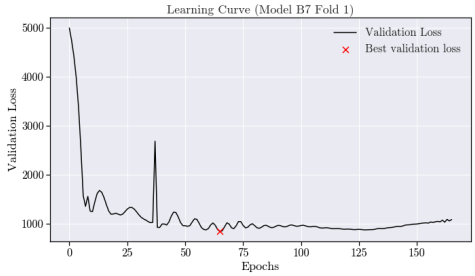
(d) Model A7 Fold 4



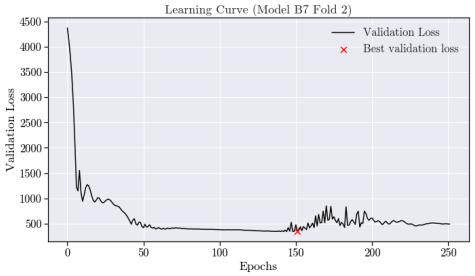
(e) Model A7 Fold 5

Figure A.13: Learning curves for Model A7 across 5 folds.

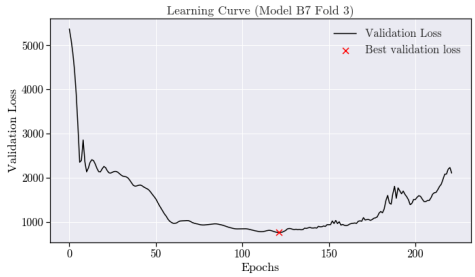
Model B7



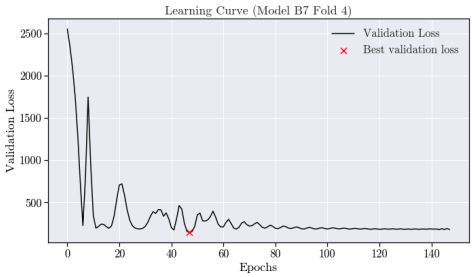
(a) Model B7 Fold 1



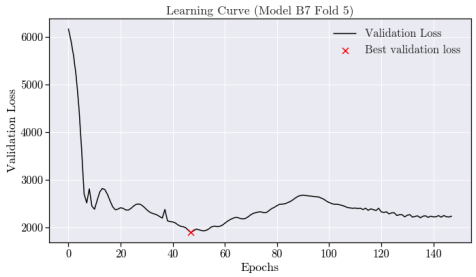
(b) Model B7 Fold 2



(c) Model B7 Fold 3



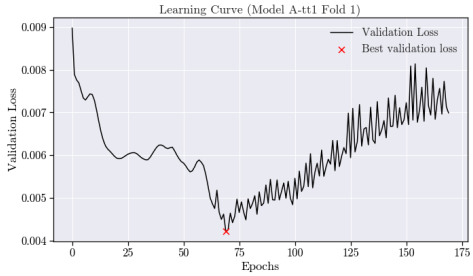
(d) Model B7 Fold 4



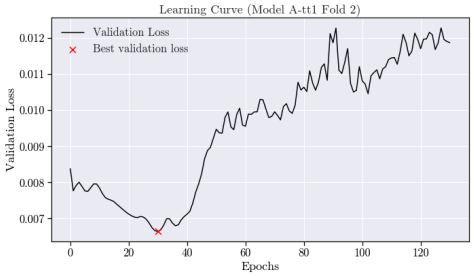
(e) Model B7 Fold 5

Figure A.14: Learning curves for Model B7 across 5 folds.

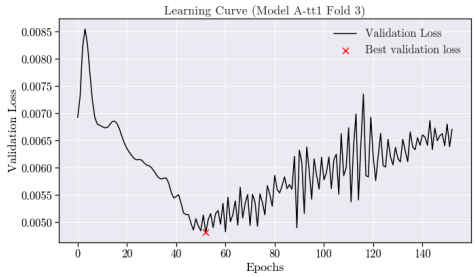
Model A-tt1



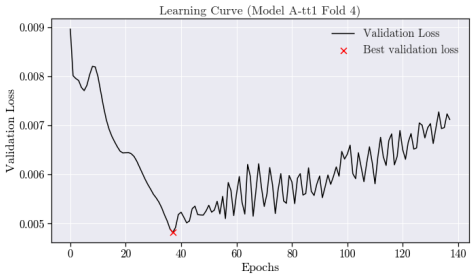
(a) Model A-tt1 Fold 1



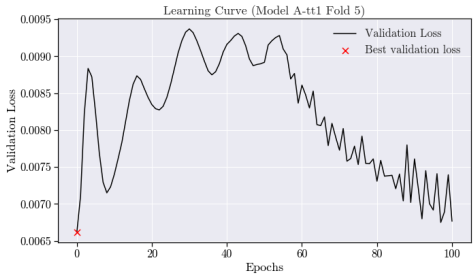
(b) Model A-tt1 Fold 2



(c) Model A-tt1 Fold 3



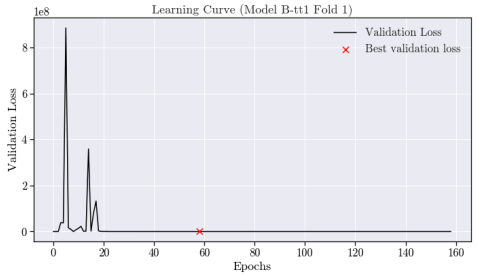
(d) Model A-tt1 Fold 4



(e) Model A-tt1 Fold 5

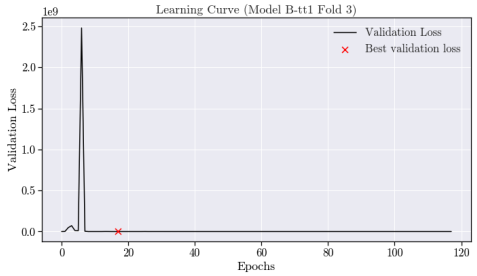
Figure A.15: Learning curves for Model A-tt1 across 5 folds.

Model B-tt1

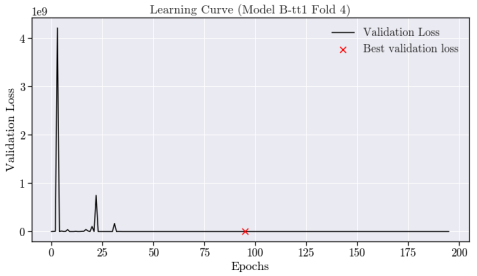


(a) Model B-tt1 Fold 1

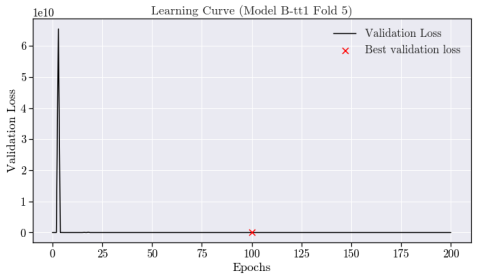
(b) Model B-tt1 Fold 2



(c) Model B-tt1 Fold 3



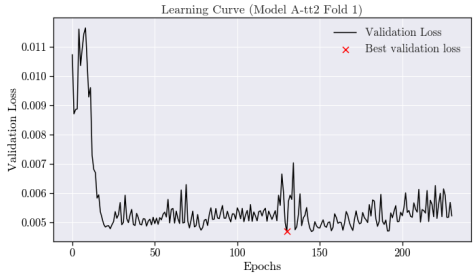
(d) Model B-tt1 Fold 4



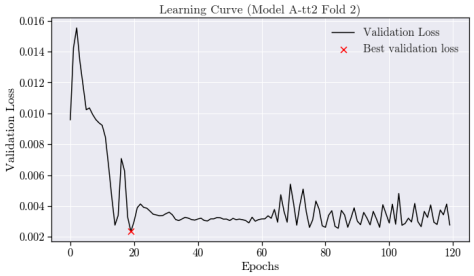
(e) Model B-tt1 Fold 5

Figure A.16: Learning curves for Model B-tt1 across 5 folds.

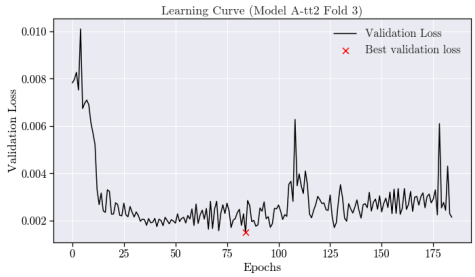
Model A-tt2



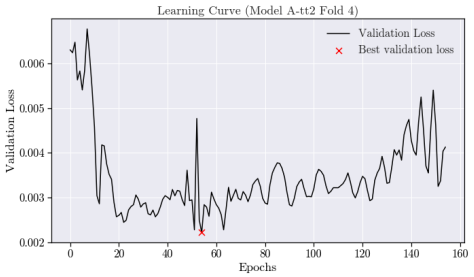
(a) Model A-tt2 Fold 1



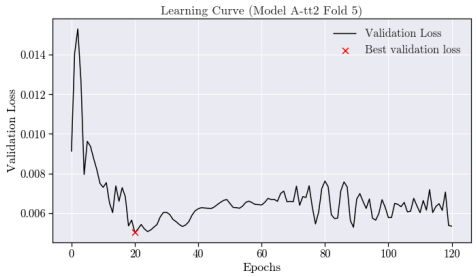
(b) Model A-tt2 Fold 2



(c) Model A-tt2 Fold 3



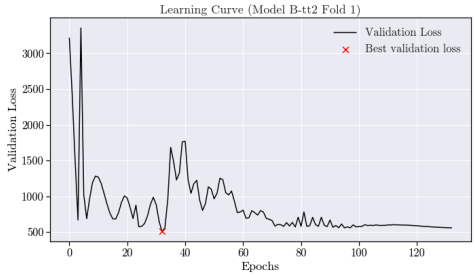
(d) Model A-tt2 Fold 4



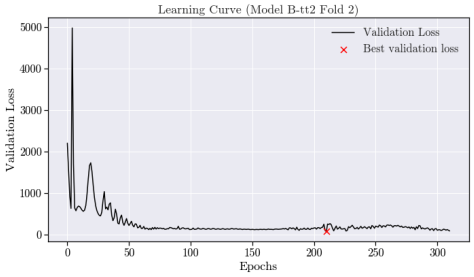
(e) Model A-tt2 Fold 5

Figure A.17: Learning curves for Model A-tt2 across 5 folds.

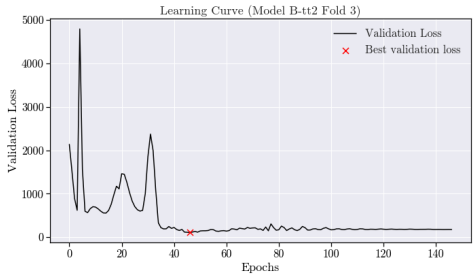
Model B-tt2



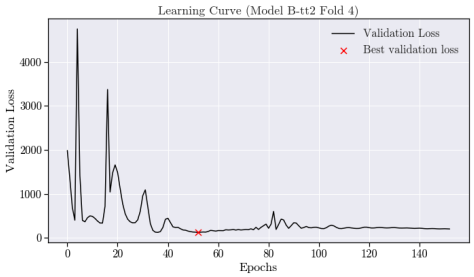
(a) Model B-tt2 Fold 1



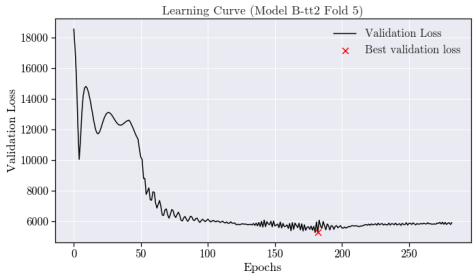
(b) Model B-tt2 Fold 2



(c) Model B-tt2 Fold 3



(d) Model B-tt2 Fold 4



(e) Model B-tt2 Fold 5

Figure A.18: Learning curves for Model B-tt2 across 5 folds.

B

Appendix B: Hyperparameters

Hyperparameter	Model A1-NN	Model B1-NN
Learning rate	0.001	0.01
λ	1e-7	1e-7
Number of hidden nodes	64	32
Number of hidden layers	16	8

(a) Neural Networks A1, B1

Hyperparameter	Model A1-RF	Model B1-RF
Number of estimators	200	100
Maximum depth	20	10
Minimum samples per split	2	2
Maximum features per split	sqrt	sqrt

(b) Random Forests A1, B1

Table B.1: Hyperparameters for models A1 and B1.

Hyperparameter	Model A2-NN	Model B2-NN
Learning rate	0.001	0.001
λ	1e-6	1e-8
Number of hidden nodes	32	16
Number of hidden layers	16	16

(c) Neural Networks A2, B2

Hyperparameter	Model A2-RF	Model B2-RF
Number of estimators	100	100
Maximum depth	20	None[1]
Minimum samples per split	5	5
Maximum features per split	sqrt	sqrt

(d) Random Forests A2, B2

Table B.2: Hyperparameters for models A2 and B2.

[1] If maximum depth is set to None: nodes are expanded until all leaves are pure or until all leaves contain less samples than the minimum number of samples per split.

Hyperparameter	Model A3-NN	Model B3-NN
Learning rate	0.001	0.01
λ	1e-6	1e-6
Number of hidden nodes	64	16
Number of hidden layers	16	8

(a) Neural Networks A3, B3

Hyperparameter	Model A3-RF	Model B3-RF
Number of estimators	200	100
Maximum depth	20	10
Minimum samples per split	2	2
Maximum features per split	sqrt	sqrt

(b) Random Forests A3, B3

Table B.3: Hyperparameters for models A3 and B3.

Hyperparameter	Model A4-NN	Model B4-NN
Learning rate	0.01	0.01
λ	1e-8	1e-7
Number of hidden nodes	32	32
Number of hidden layers	168	8

(a) Neural Networks A4, B4

Hyperparameter	Model A4-RF	Model B4-RF
Number of estimators	100	200
Maximum depth	10	10
Minimum samples per split	2	2
Maximum features per split	sqrt	sqrt

(b) Random Forests A4, B4

Table B.4: Hyperparameters for models A4 and B4.

Hyperparameter	Model A5-NN	Model B5-NN
Learning rate	0.001	0.001
λ	1e-7	1e-7
Number of hidden nodes	64	16
Number of hidden layers	16	16

(a) Neural Networks A5, B5

Hyperparameter	Model A5-RF	Model B5-RF
Number of estimators	300	300
Maximum depth	10	10
Minimum samples per split	2	2
Maximum features per split	sqrt	sqrt

(b) Random Forests A5, B5

Table B.5: Hyperparameters for models A5 and B5.

Hyperparameter	Model A6-NN	Model B6-NN
Learning rate	0.01	0.01
λ	1e-6	1e-8
Number of hidden nodes	16	64
Number of hidden layers	16	8

(a) Neural Networks A6, B6

Hyperparameter	Model A6-RF	Model B6-RF
Number of estimators	300	200
Maximum depth	10	None
Minimum samples per split	2	2
Maximum features per split	sqrt	sqrt

(b) Random Forests A6, B6

Table B.6: Hyperparameters for models A6 and B6.

Hyperparameter	Model A7-NN	Model B7-NN
Learning rate	0.001	0.01
λ	1e-8	1e-7
Number of hidden nodes	16	32
Number of hidden layers	16	8

(a) Neural Networks A7, B7

Hyperparameter	Model A7-RF	Model B7-RF
Number of estimators	100	100
Maximum depth	None	10
Minimum samples per split	2	2
Maximum features per split	sqrt	sqrt

(b) Random Forests A7, B7

Table B.7: Hyperparameters for models A7 and B7.

Hyperparameter	Model A-tt1-NN	Model B-tt1-NN
Learning rate	0.001	0.1
λ	1e-7	1e-8
Number of hidden nodes	16	64
Number of hidden layers	16	8

(a) Neural Networks A-tt1, B-tt1

Hyperparameter	Model A-tt1-RF	Model B-tt1-RF
Number of estimators	300	300
Maximum depth	20	10
Minimum samples per split	2	10
Maximum features per split	sqrt	sqrt

(b) Random Forests A-tt1, B-tt1

Table B.8: Hyperparameters for models A-tt1 and B-tt1.

Hyperparameter	Model A-tt2-NN	Model B-tt2-NN
Learning rate	0.001	0.01
λ	1e-8	1e-6
Number of hidden nodes	32	64
Number of hidden layers	32	8

(a) Neural Networks A-tt2, B-tt2

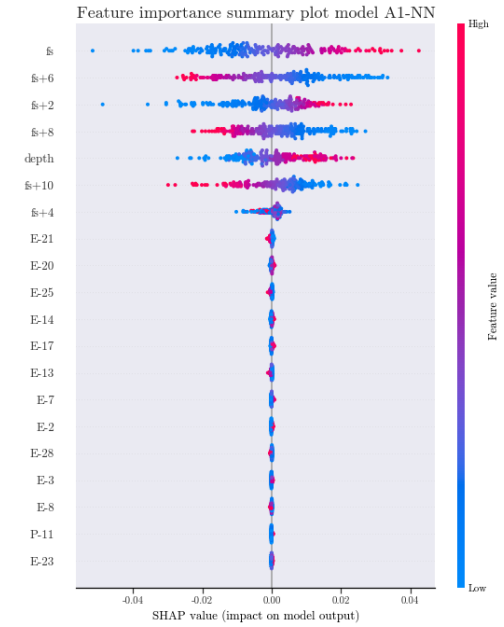
Hyperparameter	Model A-tt2-RF	Model B-tt2-RF
Number of estimators	200	200
Maximum depth	10	20
Minimum samples per split	5	2
Maximum features per split	sqrt	sqrt

(b) Random Forests A-tt2, B-tt2

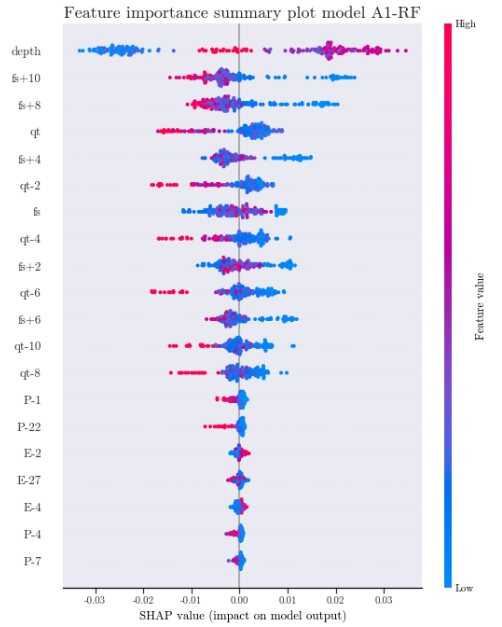
Table B.9: Hyperparameters for models A-tt2 and B-tt2.

C

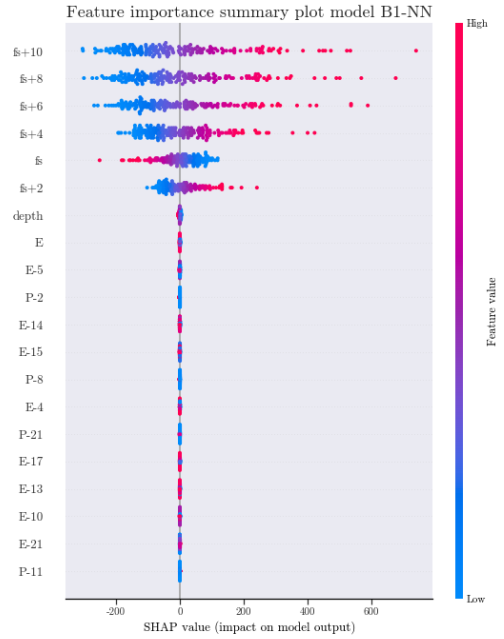
Appendix C: SHAP Summary Plots



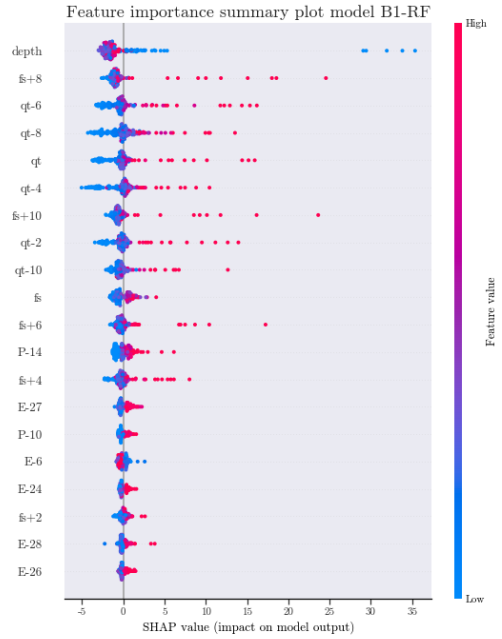
(a) Model A1-NN.



(b) Model A1-RF.

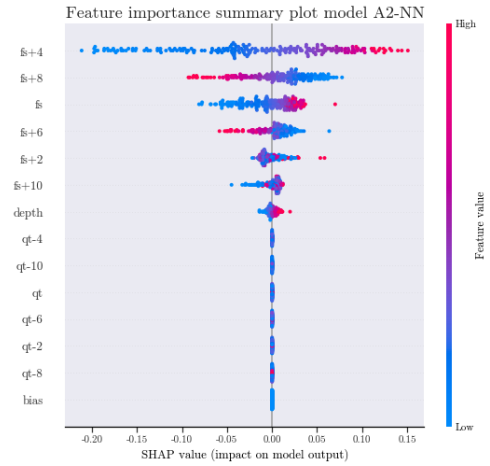


(c) Model B1-NN.

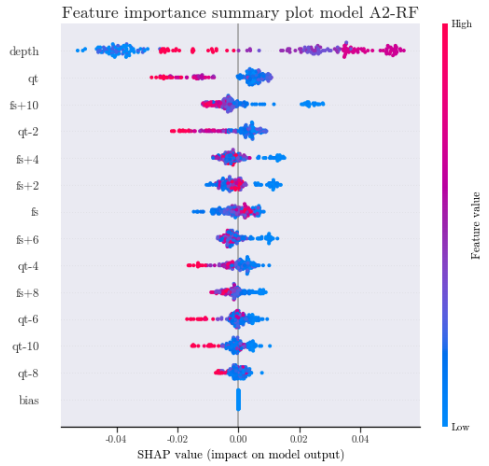


(d) Model B1-RF.

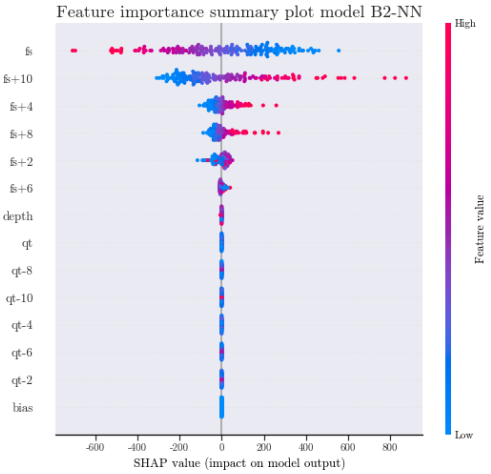
Figure C.1: SHAP summary plots of models A1 and B1.



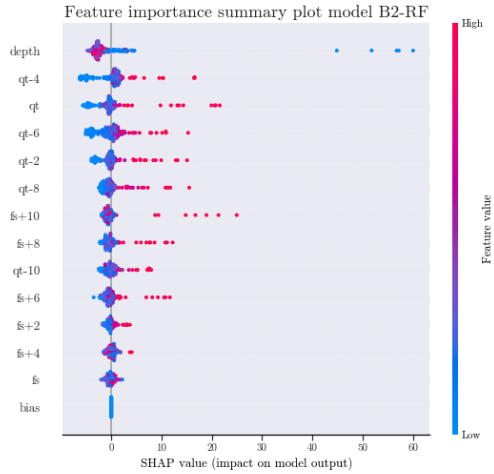
(a) Model A2-NN.



(b) Model A2-RF.

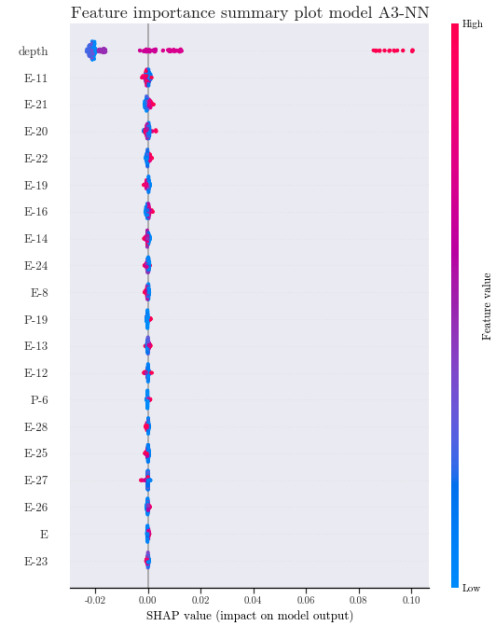


(c) Model B2-NN.

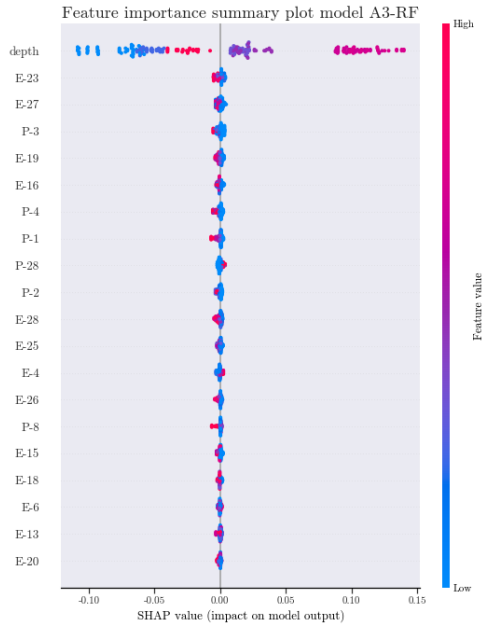


(d) Model B2-RF.

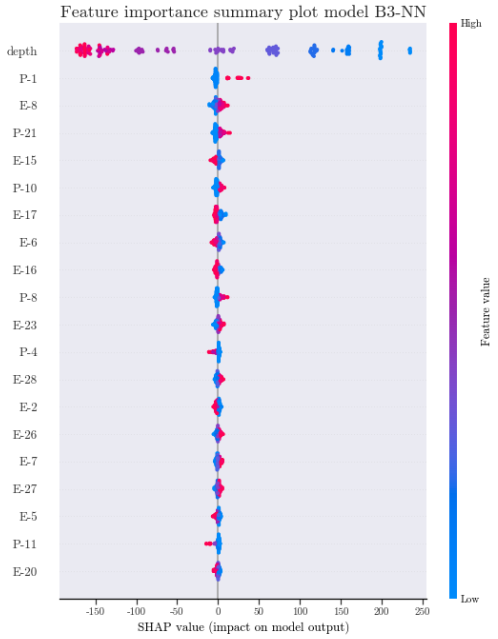
Figure C.2: SHAP summary plots of models A2 and B2.



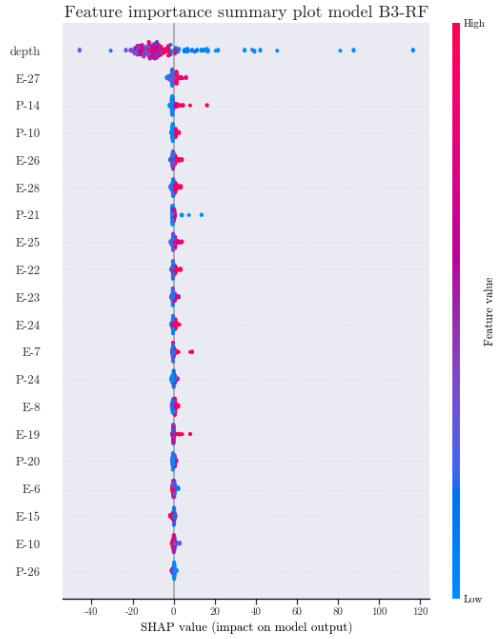
(a) Model A3-NN.



(b) Model A3-RF.

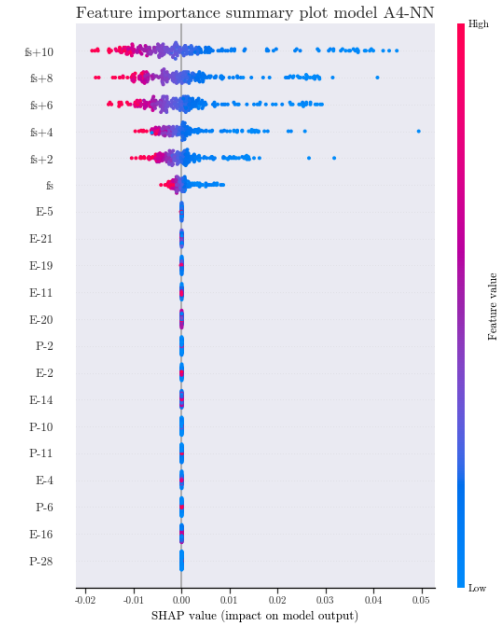


(c) Model B3-NN.

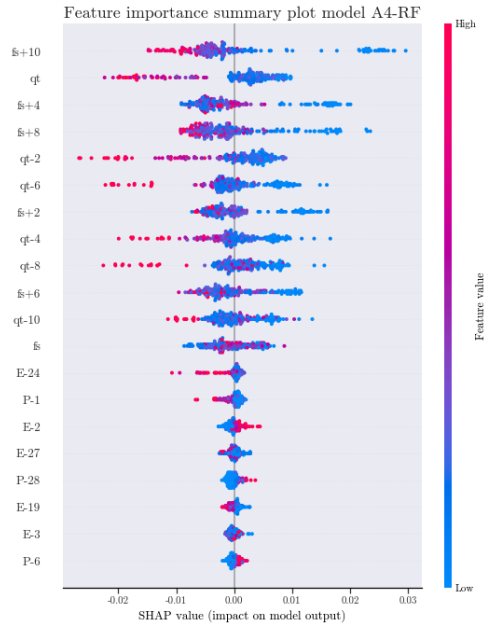


(d) Model B3-RF.

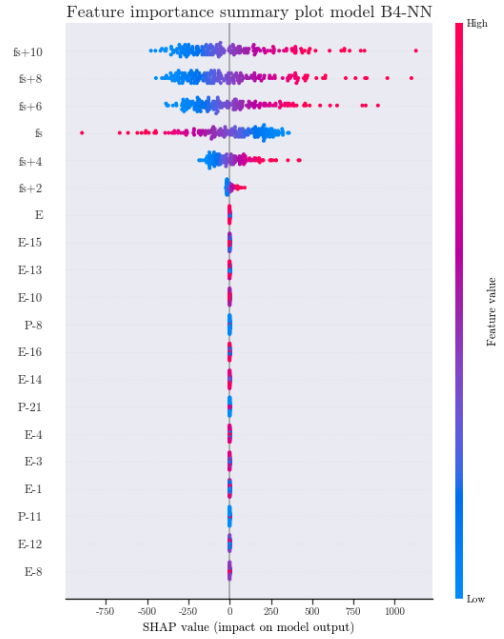
Figure C.3: SHAP summary plots of models A3 and B3.



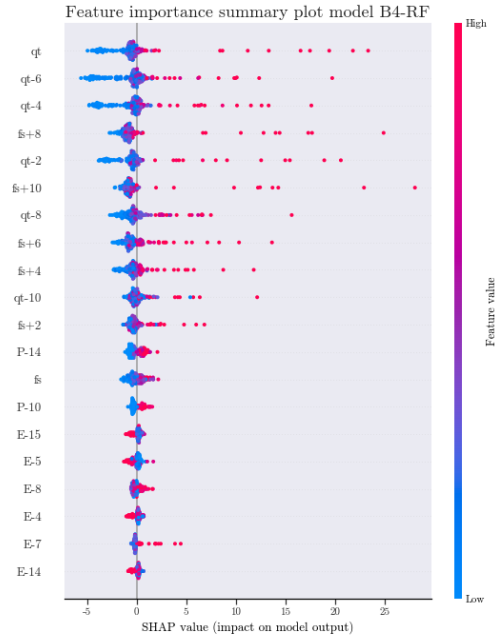
(a) Model A4-NN.



(b) Model A4-RF.

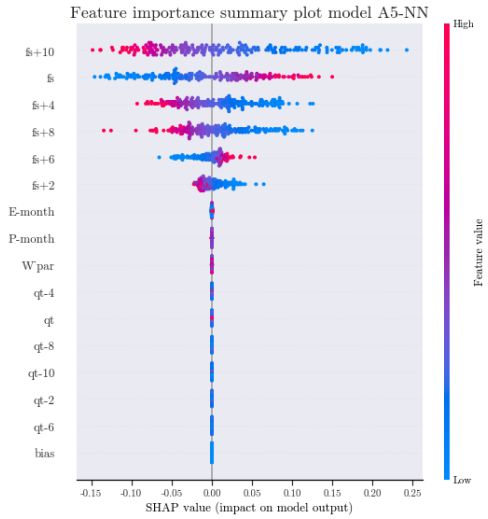


(c) Model B4-NN.

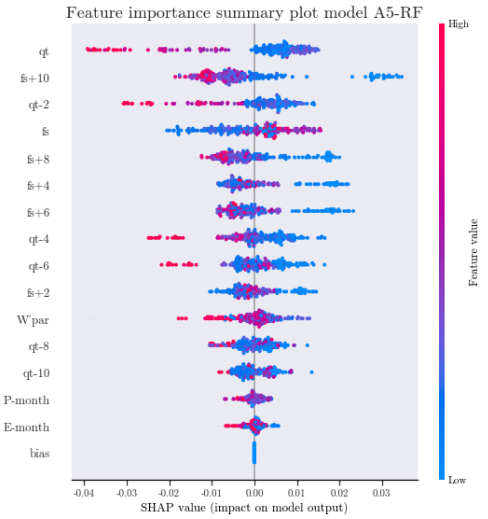


(d) Model B4-RF.

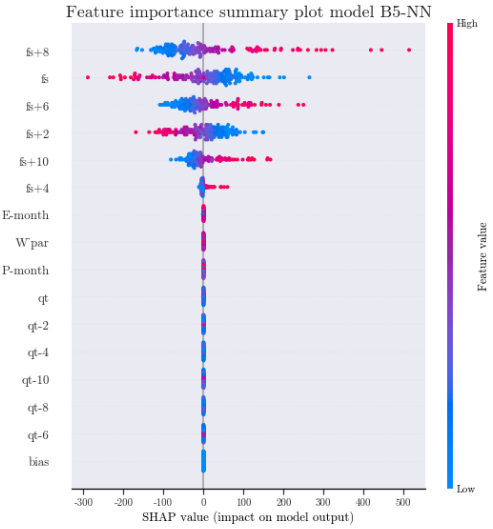
Figure C.4: SHAP summary plots of models A4 and B4.



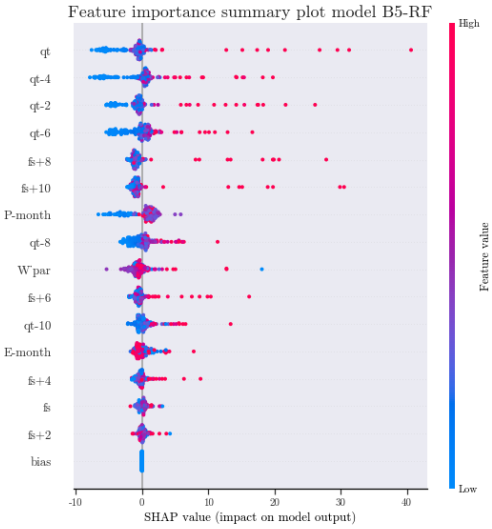
(a) Model A5-NN.



(b) Model A5-RF.

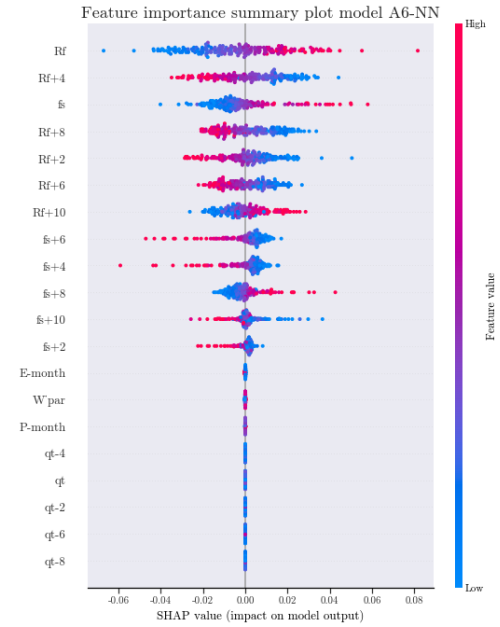


(c) Model B5-NN.

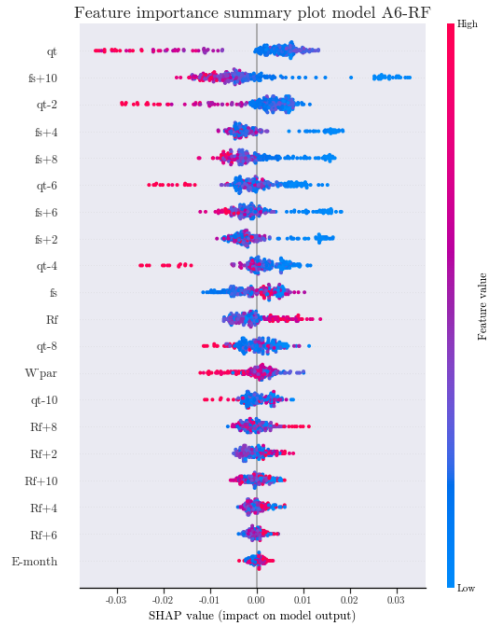


(d) Model B5-RF.

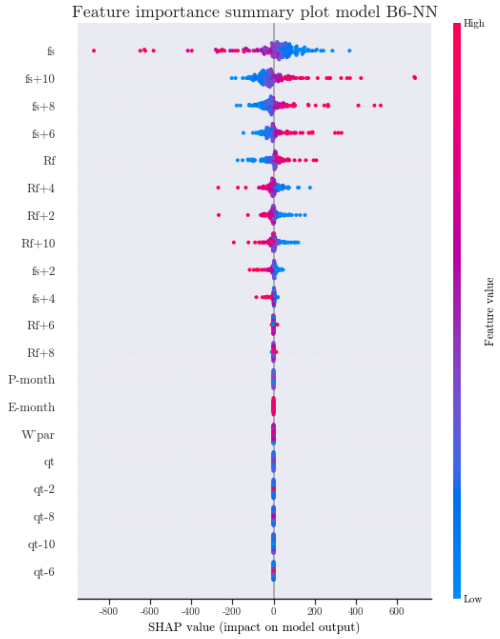
Figure C.5: SHAP summary plots of models A5 and B5.



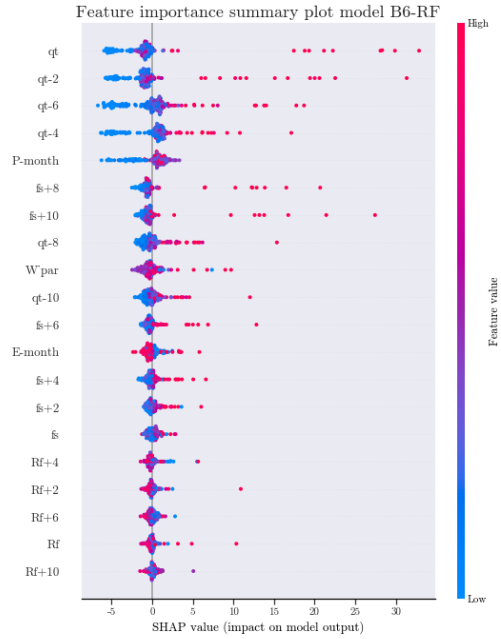
(a) Model A6-NN.



(b) Model A6-RF.

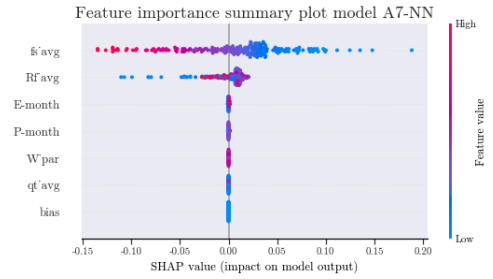


(c) Model B6-NN.

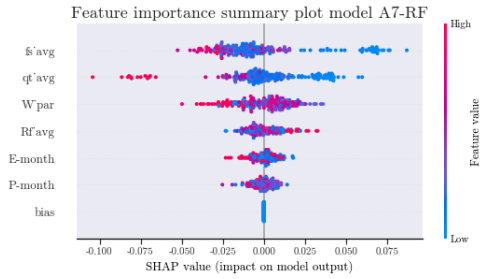


(d) Model B6-RF.

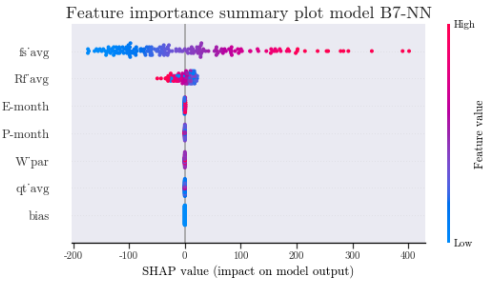
Figure C.6: SHAP summary plots of models A6 and B6.



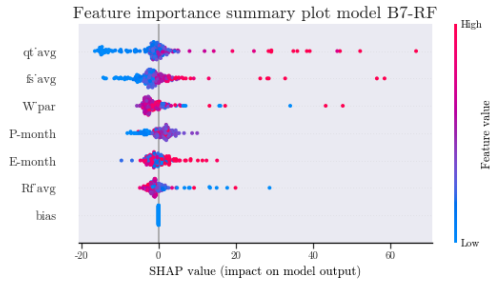
(a) Model A7-NN.



(b) Model A7-RF.



(c) Model B7-NN.



(d) Model B7-RF.

Figure C.7: SHAP summary plots of models A7 and B7.

D

Appendix D: Dependence Plots

Model A2-NN

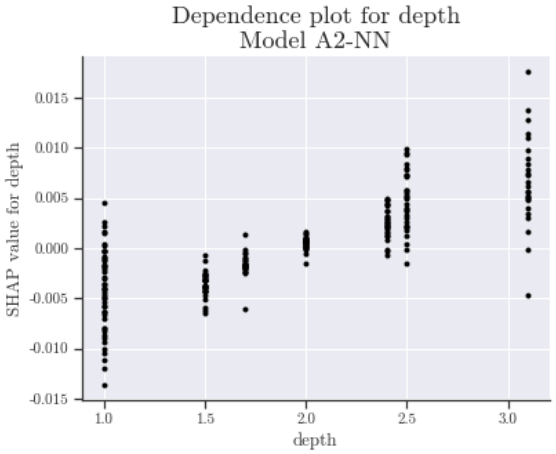
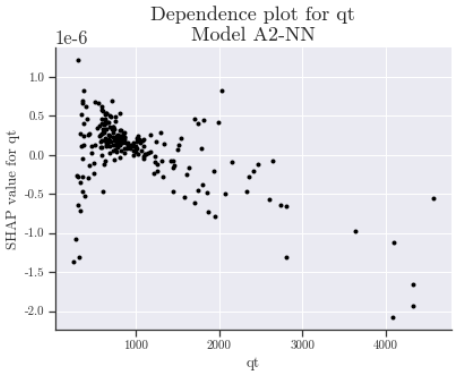
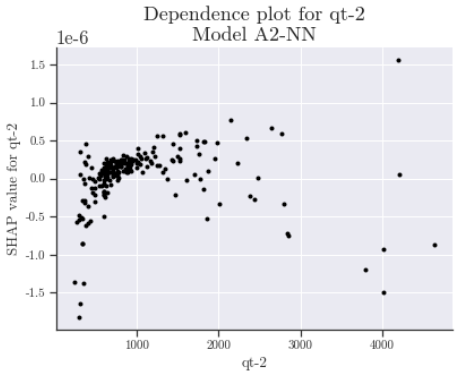


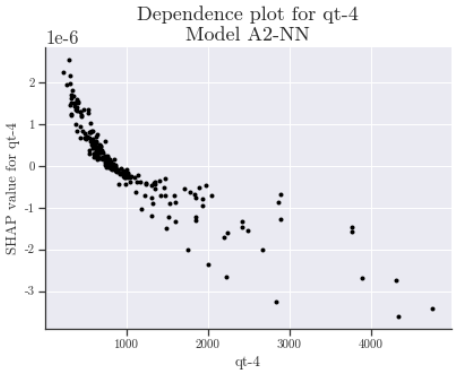
Figure D.1: Dependence plot for depth for model A2-NN.



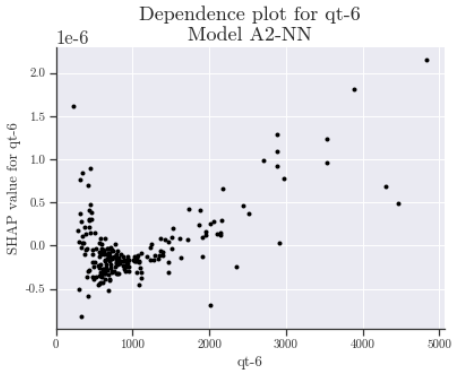
(a) q_t



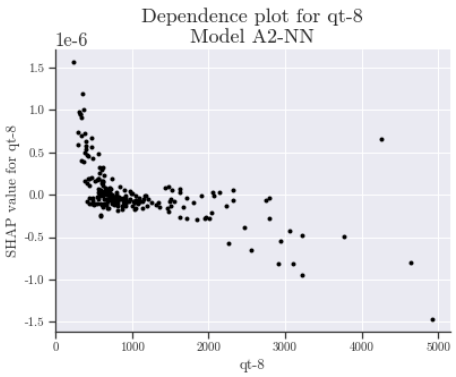
(b) $q_{t,-2}$



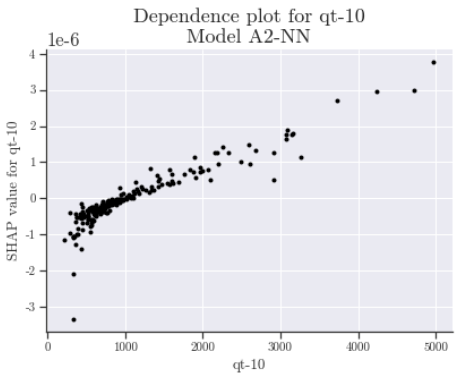
(c) $q_{t,-4}$



(d) $q_{t,-6}$

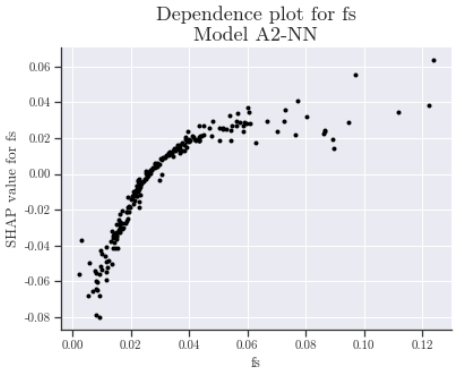


(e) $q_{t,-8}$

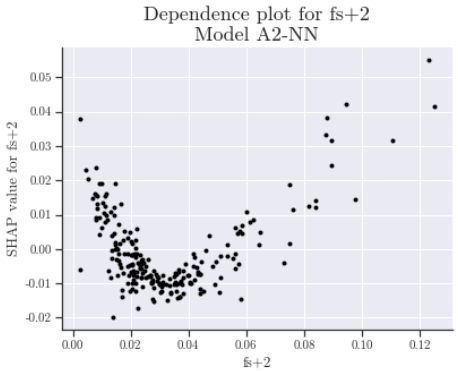


(f) $q_{t,-10}$

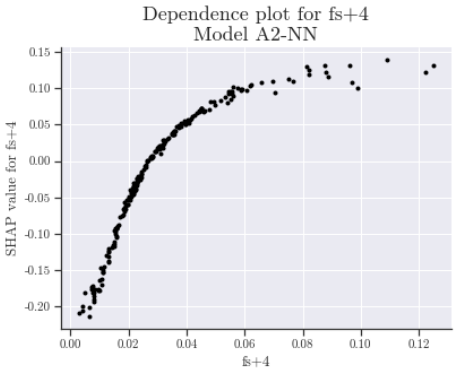
Figure D.2: Dependence plots for q_t for model A2-NN.



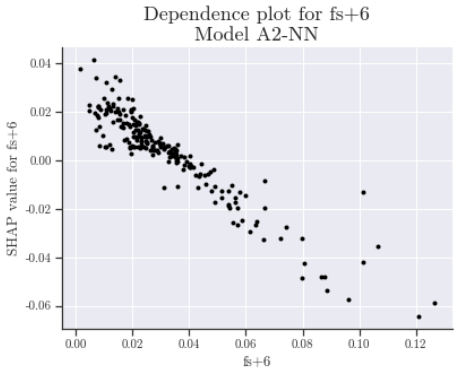
(a) f_s



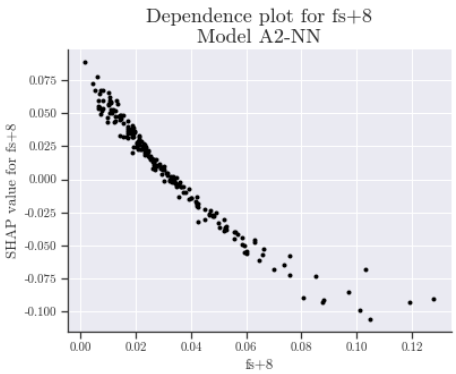
(b) $f_{s,+2}$



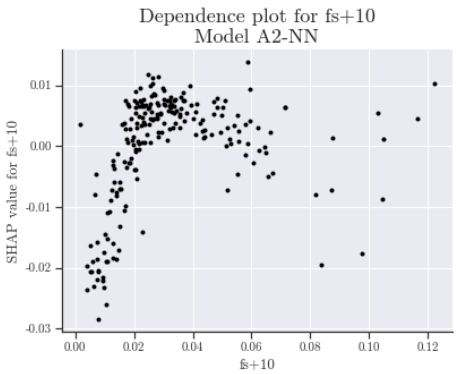
(c) $f_{s,+4}$



(d) $f_{s,+6}$



(e) $f_{s,+8}$



(f) $f_{s,+10}$

Figure D.3: Dependence plots for f_s for model A2-NN.

Model A2-RF

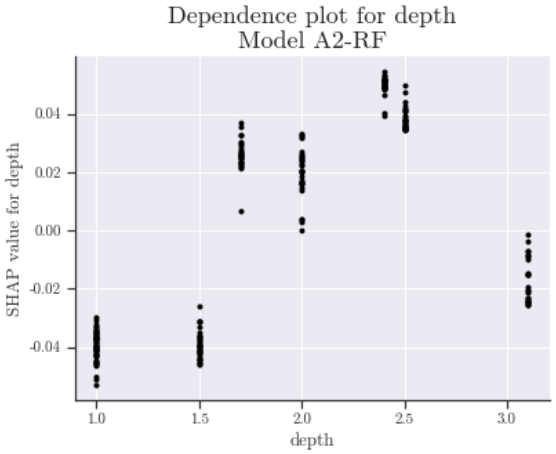
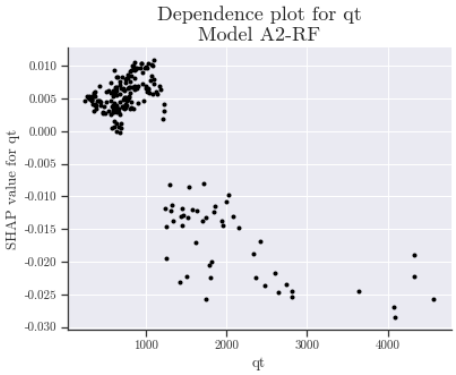
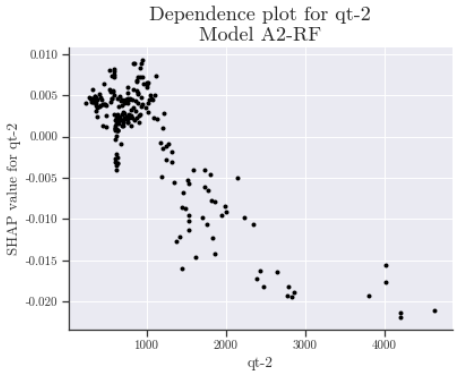


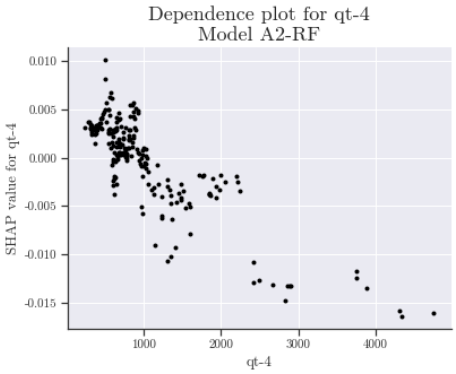
Figure D.4: Dependence plot for depth for model A2-RF.



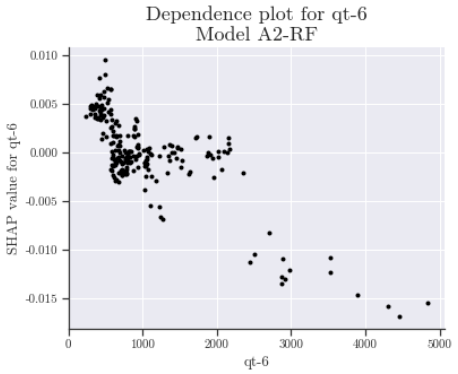
(a) q_t



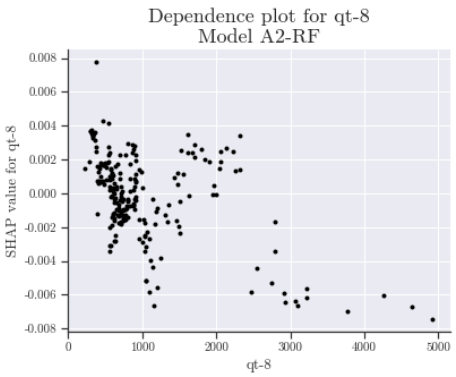
(b) $q_{t,-2}$



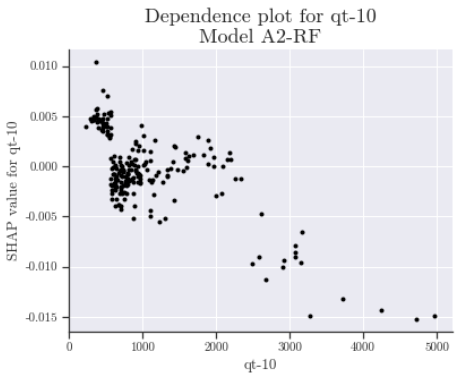
(c) $q_{t,-4}$



(d) $q_{t,-6}$

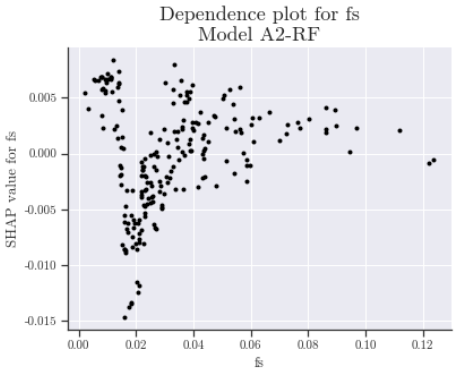


(e) $q_{t,-8}$

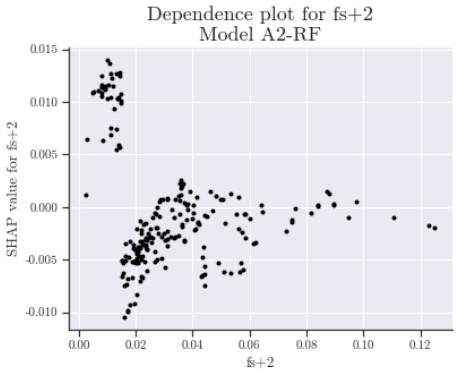


(f) $q_{t,-10}$

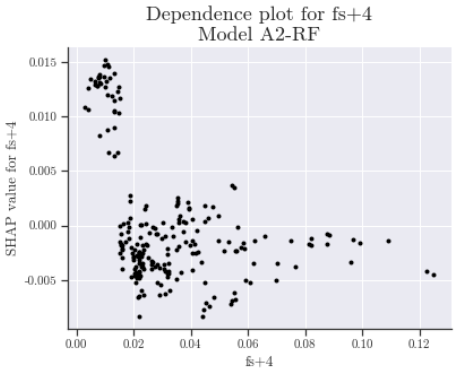
Figure D.5: Dependence plots for q_t for model A2-RF.



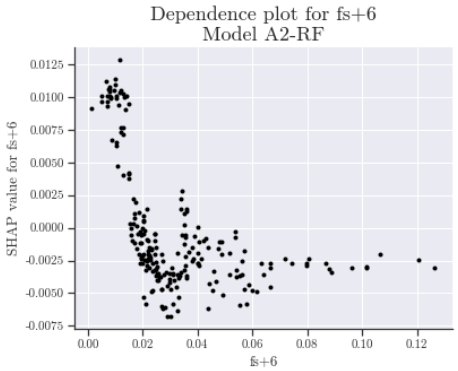
(a) f_s



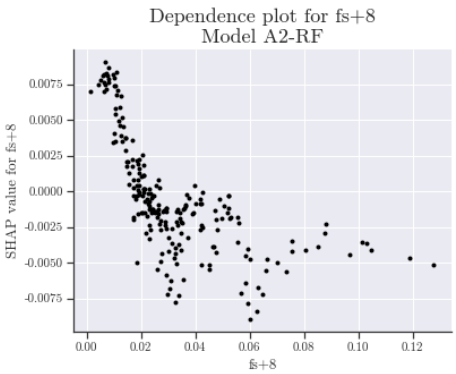
(b) $f_{s,+2}$



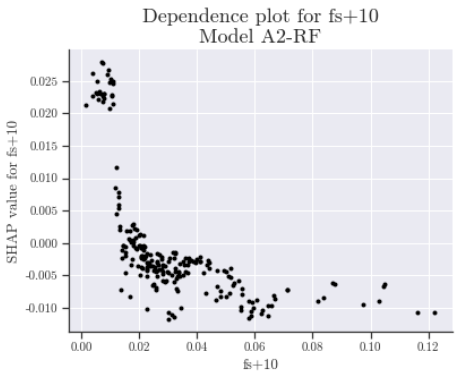
(c) $f_{s,+4}$



(d) $f_{s,+6}$



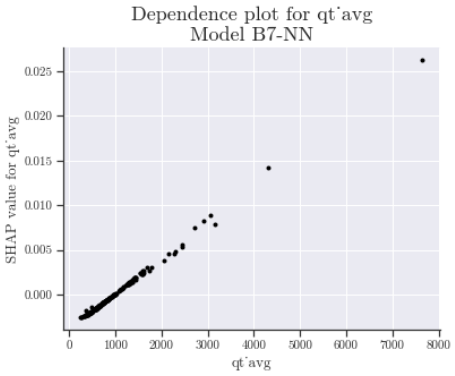
(e) $f_{s,+8}$



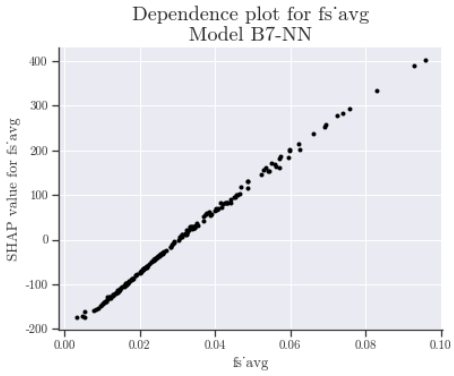
(f) $f_{s,+10}$

Figure D.6: Dependence plots for f_s for model A2-RF.

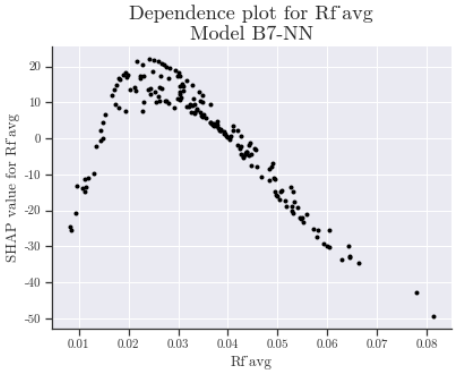
Model B7-NN



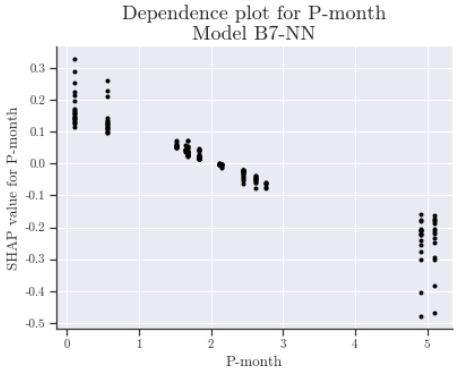
(a) $q_{t,avg}$



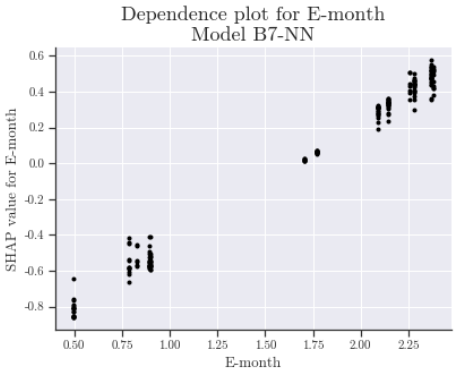
(b) $f_{s,avg}$



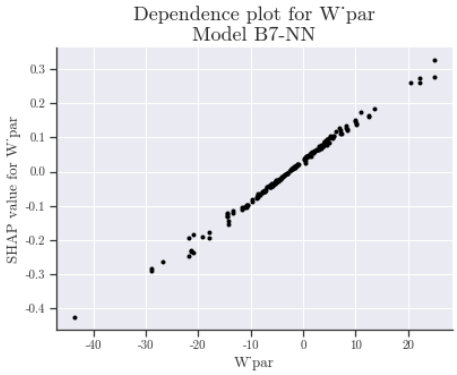
(c) $R_{f,avg}$



(d) P_{month}



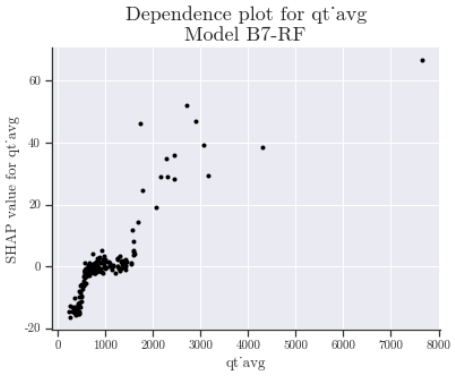
(e) E_{month}



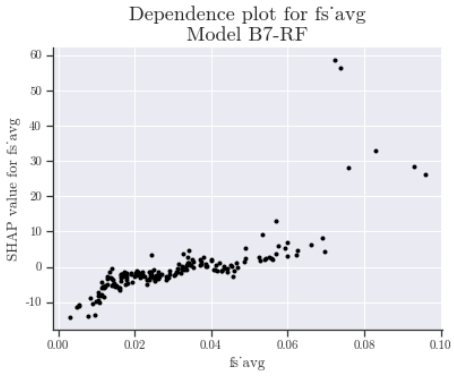
(f) W

Figure D.7: Dependence plots for all features for model B7-NN.

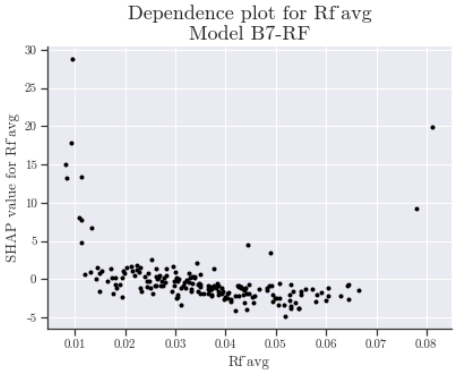
Model B7-RF



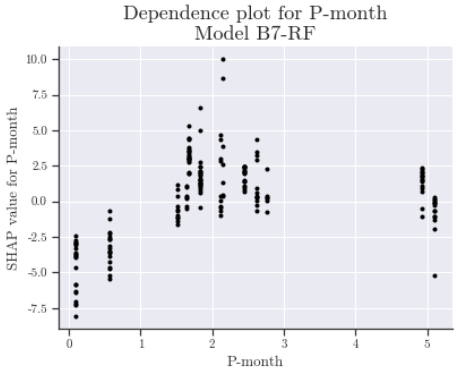
(a) $q_{t,avg}$



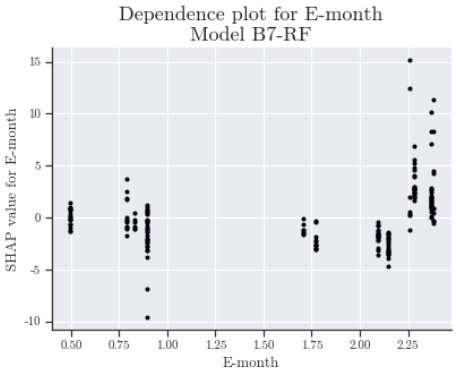
(b) $f_{s,avg}$



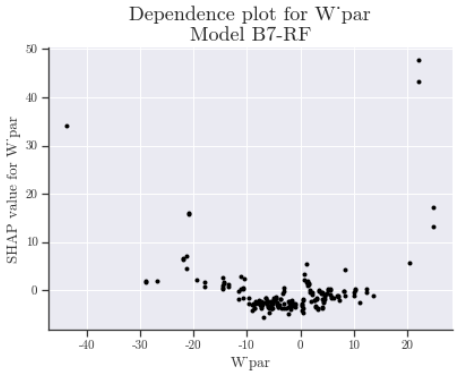
(c) $R_{f,avg}$



(d) P_{month}



(e) E_{month}



(f) W

Figure D.8: Dependence plots for all features for model B7-RF.