ARCADA

Acid Sulfate Soils Classification and Prediction from Environmental Covariates using Extreme Learning Machines

Tamirat Atsemegiorgis

Master's Thesis
Master of Engineering - Big Data Analytics
May 31, 2023

MASTER'S THESIS	
Arcada University of Applied Sciences	
Degree Programme:	Master of Engineering - Big Data Analytics
Identification number:	9203
Author:	Tamirat Atsemegiorgis
Title:	Acid Sulfate Soils Classification and Prediction from Environmental Covariates using Extreme Learning Machines
Supervisor (Arcada):	Anton Akusok and Leonardo Espinosa-Leal
Commissioned by:	

Abstract:

Acid Sulfate Soil (ASS) is a hazardous soil type primarily resulting from naturally occurring phenomena. It is a sulfate-content sediment commonly found in coastal areas around the globe and, with a notable concentration in Europe's Baltic region, highest in Finland. ASS poses a significant threat to the environment and society due to the potential to create soil acidification and leaching of heavy metals into the water ground. Therefore, it has become a national concern, and having an accurate ASS map for the nation is crucial for effective environmental management and land use planning. These days, Machine Learning (ML) methods are widely adopted for classifying soil into ASS or non-ASS types. The research aim is to explore the use of the Extreme Learning Machine (ELM) in an acid-sulfate soil classification task. The research database comes from Finland's west coast region, containing point observations and environmental covariates dataset. The experimental results show the overall accuracy of ELM and Random Forest model are the same. However, ELM implementation is easy, fast, and requires minimal human intervention compared to conventional ML methods like Random Forest.

Keywords:	Extreme Machine Learning, Classification, Acid Sulfate Soil, Environmental Covariate Map
Number of pages:	48
Language:	English
Date of acceptance:	

CONTENTS

FIGURES

TABLES

ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
AS	Acid Sulfate
ASS	Acid Sulfate Soil
AUC_ROC	Area Under Curve _ Receiver Operating Characteristics
Corine	Coordination of information on the environment
CRS	Coordinate Reference System
CV	Cross Validation
DSM	Digital Soil Mapping
DT	Decision Tree
ELM	Extreme Learning Machine
FN	False Negative
FP	False Positive
GTK	Geographic Survey of Finland
ML	Machine Learning
MLP	Multilayer Perceptron
non-ASS	non Acid Sulfate Soil
n_Job	Number of Jobs
n_estimators	Number of Estimators
QGIS	Quantum Geographic Information System
RF	Random Forest
RFECV	Recursive Feature Elimination Cross Validation
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPI	Topographic Position Index
TRI	Topographic Ruggedness Index
TWI	Topographic Wetness Index

FOREWORD

My special thanks and gratitude goes to Dr. Kaj-Mikael Björk, Head of the Department of Business Management and Analytics at Arcada, who allowed me to participate in the research project, and to my supervisors, Anton Akusok (Ph.D.) and Leonardo Espinosa-Leal (Ph.D.), for their excellent guidance and support during the process. For me, it is a great owner to work and study with the most experienced researchers and talented lecturers. Even though working on soil science and digital mapping was challenging initially because of my background, I turned it into a learning opportunity for my future career.

I also want to thank the Arcada University of Applied Science and its staff for allowing me to study a master's program in Big Data Analytics and work as a researcher.

Finally, I thank my family and friends for encouraging me to pursue my goal, and for being there throughout my journey. I would also like to thank you, my reader, and enjoy your reading.

1 INTRODUCTION

Acid Sulfate Soil (ASS) causes significant environmental challenges in Finland.This master's thesis study, based on the west coast area of Finland, focuses on classifying and predicting soil as either ASS or non-ASS using soils sample dataset and environmental covariate layers .

1.1 Background

It is a common understanding the soil of our biosphere houses various natural resources needed for life on this planet. For humans, soil is a means of agricultural production and a source of raw materials required to build infrastructures. Without it, it is impossible to think of constrictions of roads, bridges, buildings, dams, landscapes, and other human civilization symbols and technologies. However, the soil being a means of subsistence for living creatures and a reason for civilization, it is also a finite resource that needs proper attention regarding its health and usage (Eash et al. (2015), Sarangi et al. (2022)). Misuse of limited soil resources and lack of adequate prevention mechanisms has consequences on the environment and our livelihood (Eash et al. (2015), L et al. (2021)).

The well-being of humans depends on the health of the soil. Excellent and healthy soils provide healthy crops, medications, water filtration, provision of shelter, food, and clothes (Brevik et al. (2020), L et al. (2021)). The United Nations Report on Sustainable Development Goals (SDG) by 2030 indicated that healthy soil is believed to be the source of various ecosystem services, including crop production, nutrient supply, detoxification, water, nutrient retention, and maintaining biodiverse (L et al. (2021)). Again, the report stressed that healthy soil is indispensable for a country's sustainable growth and development. Degradation and depravations of soil adversely affect the availability of food and shelter and the provision of natural resources needed to construct infrastructure and produce goods and services. Hence, a deep understanding of our ecosystem's current soil structure and environmental condition is critical. Soil composition on the earth's surface is diverse, and knowing its property is essential to implement appropriate soil conservation and management strategies (Epie et al. (2014), L et al. (2021)).

The continued soil degradation of our planet occurs because of both man-made activities and naturally occurring phenomena. As mentioned before, ASS is a naturally occurring phenomenon causing soil acidification. ASS distribution persists in all continents (Lal (2017), Andriesse & van Mensvoort (2002), Huang et al. (2011)), the Baltic basin contains most of Europe's ASS, and Finland exhibits the highest aggregation (Epie et al. (2014), F et al. (2008), Yli-Halla et al. (1999)). According to GTK, ASS in Finland is considered an existential threat to the nation's environment, potentially disrupting the delicate balance of its ecosystem (Jaakko et al. (2022)).

The exposure of ASS to oxygen above see-level generates sulfuric acid that leads to the acidification of soils and the release of heavy metals into the waterbodies(Jaakko et al. (2022)). This phenomenon creates a toxic environment for aquatic plants and animals(Palko (1986)) and has a huge impact on fishing and agricultural production of the nation; for more, read the manuscripts (Eden et al. (1999), Joukainen & Yli-Halla (2003), Yli-Halla (2022)). Therefore, it is imperative to implement policy-based soil conservation and soil management strategies to prevent the damage caused by ASS(Ministry of Agriculture and Forestry Ministry of the Environment (2011)).

In Finland, from as early as 1950, researchers have been actively involved in mapping Acid Sulfate Soils (ASS) through a labor-intensive process of collecting soil samples from specific locations and subsequently conducting pH level analyses in laboratory settings to ascertain the presence of ASS (Ministry of Agriculture and Forestry Ministry of the Environment (2011)). However, in recent times, the use of machine learning (ML) techniques for digital mapping has emerged as a more cost-effective and streamlined approach, offering detailed soil maps. For further insights into this innovative method, please refer to the referenced manuscripts. (McBratney et al. (2003), Beucher et al. (2015), Minasny & McBratney (2016), Estévez Nuño (2020), Baltensweiler et al. (2021), Estévez et al. (2022))

1.2 Research Significance

As previously noted, acid sulfate soils (ASS) significantly influence on the environment and infrastructure we build. Regions where ASS soil is present are the potential for sulfuric acid generation and metal leaching, posing challenges to the country's socio-economic growth and citizens' livelihood (Jaakko et al. (2022)). Therefore, it is imperative to take into account the presence of ASS in all land use planning and decision-making processes.

Developing a comprehensive map of ASS distribution is crucial for implementing effective prevention and management techniques. Various efforts were made to create such a map in the past decade (Ministry of Agriculture and Forestry Ministry of the Environment (2011)). With the availability of extensive soil science and environmental data, machine learning techniques offer an alternative approach to digital ASS mapping, reducing the reliance on costly fieldwork and tedious PH measurements of soil samples (Ministry of Agriculture and Forestry Ministry of the Environment (2011)). This thesis aims to explore the performance of the Extreme Learning Machine (ELM) model in the classification of acid sulfate soils. Additionally, the research seeks to assess the comparative advantages of using ELM in contrast to conventional classification models like Random Forest (RF).

1.3 Limitation

The previous studies have typically encountered the following limitations: Firstly, traditional methods are characterized by high costs and time-intensive operations. Secondly, there is a need for more accessible experimental data. Thirdly, acquiring environmental covariates data is not straightforward, as it demands a significant amount of time and a high degree of expertise to prepare the layers for each covariate.

1.4 Research Questions

As mentioned, various ML methods were implemented to classify soil as ASS or not. The goal is to build a holistic digital ASS map for the nation, yet researchers are exploring different techniques to achieve the target. The main research topic of this thesis paper is

- How well can we classify ASS using covariate map tiles for input?
- Can Extreme Learning Machine (ELM) classify ASS correctly?

- Which features or attributes are most significant in soil classification?
- How does the ELM model perform compared with RF?

To find an answer to the above research questions, sample point observations from the west coast of Finland (prepared by GTK) and 13 environmental covariates layers were prepared using remote sensing datasets. The ELM model was explored, and its results were compared with the RF model.

Apart from the Introduction, this master's thesis is organized into six sections. The following section is called Related Works; it summarizes the previous research works done by researchers on digital soil mapping. The third section, Research and Database, prepares point observations and environmental covariate datasets. The fourth section is the Research Methodology about the machine learning models and map tiles concepts. The fifth section is the Project Framework, which focuses on the model development process, data preparation, parameter tuning, model selection, model training, and model evaluation. The project experiment runs on Jupyter Notebook and uses cutting-edge libraries GeoPandas, PySpark (a python library engine for large-scale parallel data processing)(Li et al. (2020)), and others. The sixth section is the Experimental Result, which presents the comparative results and evaluations of the proposed models, and the last section presents the Conclusions of the project works.

2 RELATED WORK

Artificial Intelligence (AI) has been known as a field of study in computer science since the 1950s (Domingos (2012)). It is a broad term focused on software development, enabling a computer system to exhibit human intelligence and behaviors. We, humans, try to learn new things by adapting to the world we live in and acting accordingly. Likewise, AI focuses on creating "intelligent devices" that imitate human characteristics in perceiving their environment and performing tasks without human intervention (Domingos (2012), Bernard M (2019), Espinosa-Leal et al. (2020)).

Author Samuel, a pioneer in AI and Machine Learning (ML), defines ML as "A Field of study that gives computers the ability to learn without being explicitly programmed.". That means ML is an application of AI in which a computer system is provided with a large volume of input data and a program called an algorithm to learn patterns and behavior of the data on its own. Through this learning process, the machine will be able to maximize an experience to generate a general rule to decide on a given task. Therefore, ML is about how AI systems "learn" the environment through data and be able to imitate human behavior and get the capability to analyze the task and make an informed decision. Today, ML is widely used in healthcare, engineering, telecommunication technology, and data-driven research; some use cases include automatic recommendation systems, fraud detection, search engines, stock marketing, social media applications, DNA sequencing, and many more (Bernard M (2019), Andreas C.Müller (2016)).

In research conducted by Virginia Estévez (Estévez Nuño (2020), Estévez et al. (2022)), datasets collected from the southeastern region of Finland, specifically Virolahti and its vicinity, were employed. The primary objective was to investigate the utilization of machine learning techniques in soil classification and the creation of a probability map for AS. To achieve this, the researcher utilized various methods, including Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF), and convolutional neural network (CNN). The finding revealed that both GB and RF methods demonstrated strong performance in soil categorization, outperforming SVM and generating superior AS probability maps. Notably, the model's probability map demonstrated enhanced objectivity and accuracy compared to traditional maps.

The Virolahti datasets were also subjected to analysis using the ELM (Extreme Learning Machine) model in a study conducted by Akusok in 2023 (Akusok et al. (2023)). The project used same databases as the previous one (Estévez Nuño (2020)). The findings of this paper's results show that ELM performance is comparable to alternative methods, SVM, and ensemble decision trees. The author acknowledged that a small training dataset was the limitation, and the researcher expressed his anticipation of better ELM performance by including additional training data.

This project aims to develop high performing ELM model to classify AS soil types using sample point observations and environmental covariates datasets. ELM model is a choice because it exhibits; fast learning behavior, high accuracy, and easiness to use. To alleviate small-size training restraint, 5824 rows of point observations from the west coast area of Finland were prepared, and more details about the datasets are found in the next section.

3 RESEARCH DATABASE

The database under study is a combination of two spatial datasets; a vector dataset (sample point soil observations), some examples are shown in Table1, and a raster dataset of map tiles(environmental covariates layers). A spatial dataset is a collection of observational attributes of phenomena organized in a tabular format with its unique ability to represent a geographical location worldwide. GTK is the provider of the points dataset. However, the covariates layers dataset was generated by using the QGIS tool based on remote sensing data.

Table 1. Some sample soil type from the west coast of Finland.

X	Y	class
25.768938	64.777988	ASS
25.776304	64.793496	ASS
25.784691	64.786808	ASS
25.315115	64.988732	ASS

3.1 Point Observations Dataset

As mentioned before, the points spatial dataset used in the experiment was provided by GTK, and Table1 presents geographical location information: longitudes (X), latitudes (Y), and a binary variable "class" which contain two classes of soil types, acid sulfate soil and non-acid sulfate soil(ASS and non-ASS).The catagorization of soil types were condacted in the laboratory based on specific criterias (PH level and others) (Estévez Nuño (2020), Estévez et al. (2022)). To visualize the sample points observations, Fig. 1 is plotted to show the heatmap distribution of soil types for the training dataset of the northwest region. The red shade area in the map is the ASS area, and its coverage area is vast.

3.2 Environmental Covariate Layers

As discussed, the covariate layers of map tiles were extracted using remote sensing data, and for further details, please consult the reference (Estévez et al. (2022), Estévez Nuño (2020)). The tiles are a single-layer grayscale image that measures a single characteristic or attribute, for instance slope, hillside, etc. To understand it better, Fig. 2 and Fig. 3 are presented to show hillshade and slope tile maps for zoom level 12 as an example.

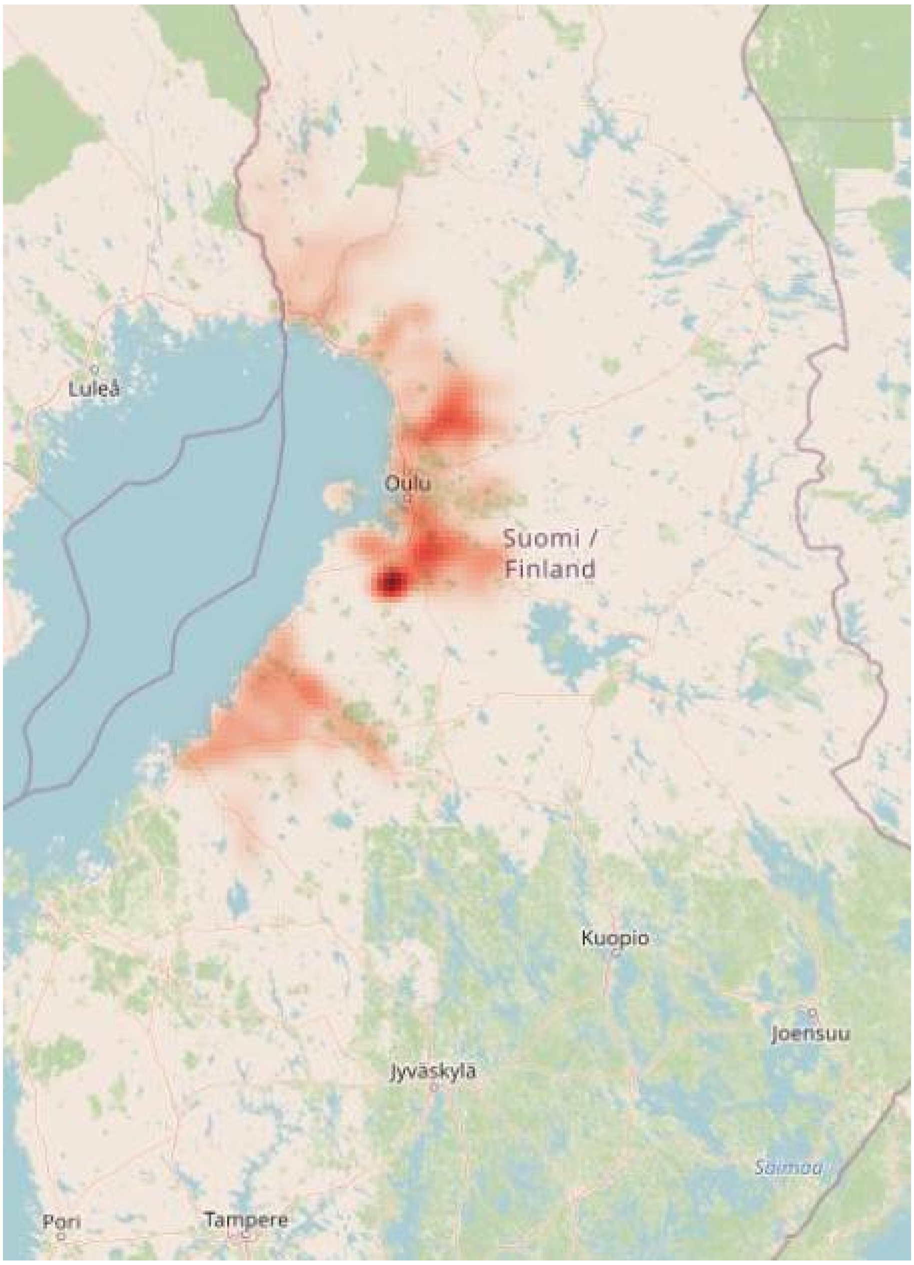


Figure 1. Heatmap of acidic soils distribution of west coast of Finland.

Similarly, 11 other environmental covariates layers were prepared using image processing techniques.

The covariates layers used in the experiment are composed of three distinct groups of layers:

1. Terrain: includes Slope, Aspect, Hillshade, Topographic Wetness Index (TWI), Topographic Position Index(TPI), Normalized Difference Vegetation Index (NDVI),

Figure 2. Hillshade tile of experimental area for zoom level 12

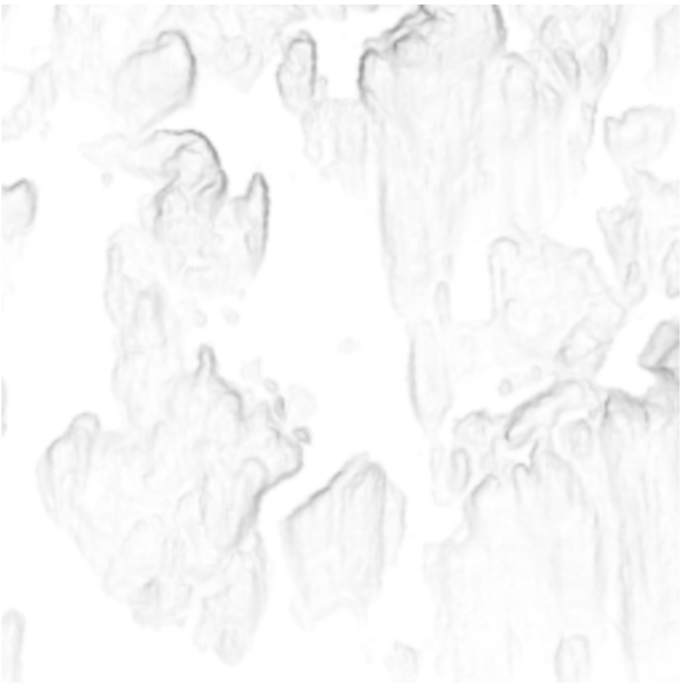


Figure 3. Slope tile of experimental area for zoom level 12

and Topographic Ruggedness Index (TRI) layers,

2. Geophysics (magnetism or electric conductivity data): includes electromagnetic real, electromagnetic imaginary and electromagnetic resistivity layers,

3. Quaternary map: 41 soil types in our case "bedrocks" and 49 different classes of land cover classes(Corine land cover) were included.

Corine stands for Coordination of Information on the environment, is a European way of land cover inventory, and has five main categories built-up areas; agricultural areas; forests and open fabrics and rocky lands; wetlands and open marshes and water areas. The Corine data contain 49 sub-categories(SYKE). TWI measures the tendency of an area to accumulate water, i.e., how likely the area is wet. An area with a higher TWI index value is more wet relative to the lower index values of the neighboring area(Deenik (2021)).TPI measures the altitude of a point against the neighboring points and hence helps to distinguish the topographic features like a hilltop, valley bottom, etc... Higher altitude point has quantified to positive TPI value and sunken points to a negative value(Čučković (2019)).

NDVI is an annual index measurement of the amount of green vegetation in the area based on the information obtained through remote sensing data. The NDVI pixel value of the dataset is 10m, and the index values range between 1 and -1 (the vegetable area is indexed positive, and the water area is indexed negative)(GISGeography (2022)). TRI measures the elevation difference between adjacent cells of the Digital Elevation Model, and the geographical heterogeneity (TRI) measurement is taken from the center cell to eight surrounding cells. The recommended classification ranges are; 0-80 level terrain surface, 81-116: nearly level surface, 117-161: slightly rugged surface, 162-239: intermediately rugged surface, 240-497: moderately rugged surface, 498-958: highly rugged surface, above 959: extremely rugged surface (Evans, Riley et al. (1999)).

In the data preparation stage, the points coordinates of the data points are extracted using the mathematical formula from the longitude and latitude pair of the sample observations. After that, the environmental covariate layers of the data point were mapped to the point coordinates of data points to extract tile coordinates points for every 13 covariates tiles maps. Finally, a database is created from those two datasets. PySpark(Python integration of Apache Spark) functions were developed to accomplish those tasks, and using it benefits speed and fault tolerance in a net shell.

The combined database consists of 5824 rows and 104 columns. Among the features, the "class" variable is a binary class of soil types (ASS, non-ASS) and is a target feature for the experiment. There are 3490 ASS and 2334 non-ASS soil types, and their frequency distribution in percentages is 60% and 40%, respectively. The database has 2 features of coordinate points (X and Y) and 13 environmental covariates layer coordinates. Among the covariant layers, features "Corine-land-cover" and "bedrock" consist of 49 and 41 distinct values, respectively. A one-hot encoding technique was used to represent those values. The encoding method adds extra 88 features: corine1 ... corine49 from Corine-land-cover and bedrock1 ... bedrock41 from the bedrock layers; hence, the number of features rose to 104.

4 RESEARCH METHODOLOGY

As discussed before, ML is an application of AI where the computer system learns to perform a task by figuring out a generalization rule from the dataset fed. It is an ever-growing field of science, and today there are thousands of ML algorithms known to be present. There are two ways of classification of algorithms: based on learning methods (including ensemble, supervised, unsupervised, semi-supervised, and reinforcement learning methods algorithms.) and second based on functional similarity(including Neuron-Network-based and tree-based methods) with some exceptions (Sullivan (2017)).

An Artificial Neural Network (ANN) based system imitates the human neural system for prediction; ELM is an excellent example and it will be discussed in the next section. However, the decision tree(tree-based) algorithm generates a model of a decision tree of a fork-like structure based on the dataset's attributes until a prediction is made for the given task.

This research experiment employs Extreme Machine Learning (ELM) using environmental covariate map tiles and compare the performance with the conventional model Random Forest.

4.1 Map Tiles

A map on a web browser is a dynamic square map composed of multiple images called Map tiles, making it easy for users to zoom and browse around the map. Google is the inventor of the Map tile system (Forrest (2023b)) developed to create a Google Maps App, and then every Maps API providers adopted the tiling technique as a standard in developing web maps . A tile is usually a 256 X 256 pixels square image with a fixed geographical area and scale. Browsing or zooming a map on the browser technically means displaying multiple image tiles very fast in a grid system, as if panning on a single image or browsing a portion of the map without loading the entire map. Zooming in and out of a map happens because of loading a new set of tiles, and each zoom level has its own sets of tiles. There are around 23 zoom levels: 0 to 22; Zoom 0 loads the entire world in a single tile; Zoom 1 contains 4 tiles; Zoom 2 contains 16 tiles, etc. Fig. 4 shows map tiles for zoom levels 0 and 1 (MapTiler (2023), Forrest (2023b,a)).

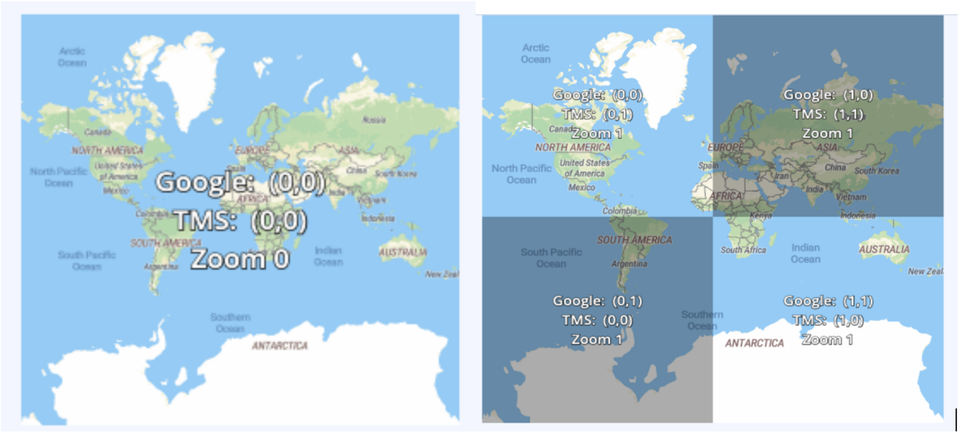


Figure 4. Tiles for zoom level 0 and 1, source (MapTiler (2023))

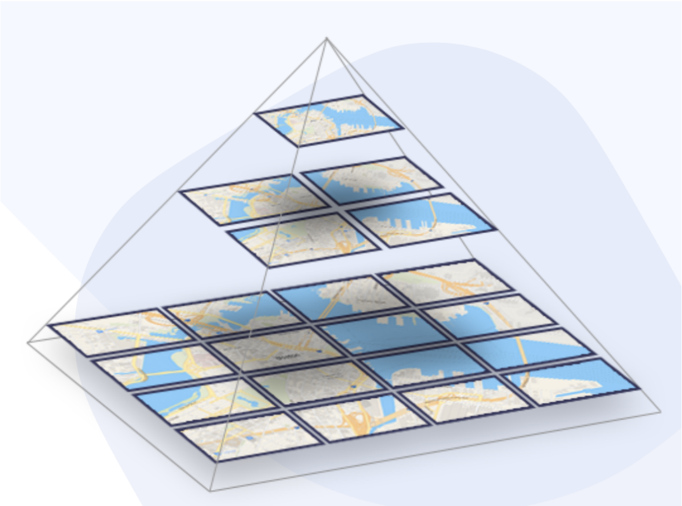


Figure 5. Map tiles for specific Zoom level at each surface of the pyramid, source (MapTiler (2023))

The tiles are arranged in pyramidal structural layers of multiple floors of zoom levels, shown in Fig. 5. Each tile has 3 assigned coordinates x,y,z, where z is the zoom level and x, y is the grid position. For instance, for zoom 1, we have (0,0), (0,1), (1,0), and (1,1) grid tiles shown in the Fig. 4, hence z/y/x in the pyramid refers to the tile coordinate or address(MapTiler (2023)).

4.2 Decision Tree and Random Forest

Before discussing Random Forest (RF), it is essential to understand the basics of decision trees (DT). DT is a building block for random forest and other tree-based ensemble models. Fig. 6 shows a random forest model for N subgroups; the blue shaded boxes in the figure represent a decision tree based on the database's subset groups. Unlike the usual

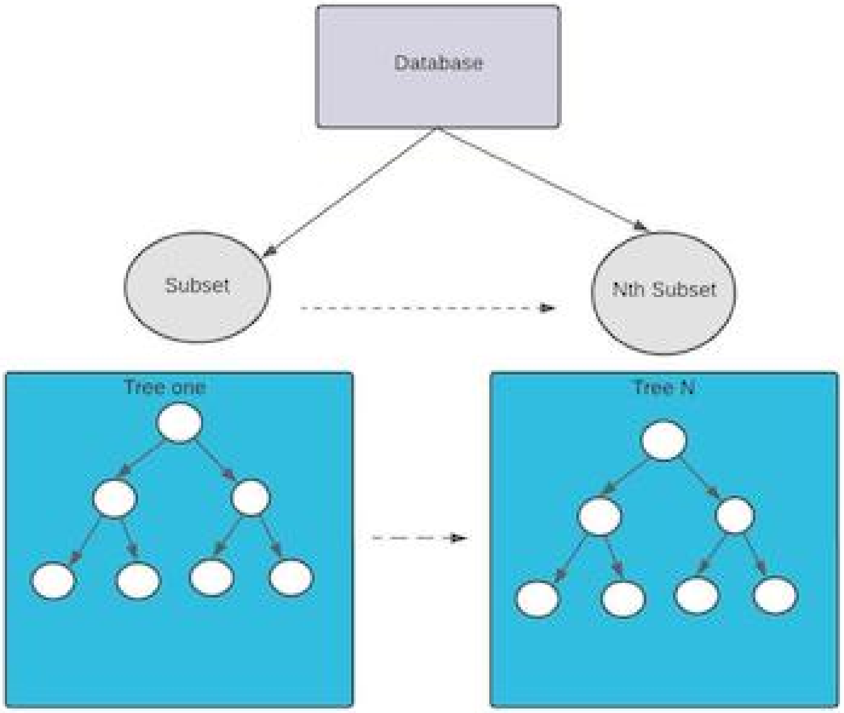


Figure 6. Random Forest model

tree, the root node is located at the top, and the leaf nodes are at the bottom, which is an upside-down tree. DT splits the node on all available features and selects the split, which results in more homogeneous sub-nodes. There are multiple algorithms used by DT to decide the best split for the given task. The most popular are Gini Impurity, Chi-Square, Information Gain used for categorical features, and Reduction in Variance for non-categorical features. For more details, refer to the page (Sharma (2023)). The algorithm tries to split the subgroups differently or equivalently; members of each subgroup are selected to be as similar as possible (Sullivan (2017), Yiu (2021)).

The decision tree algorithm works well for classification and regression types of problems; however, it's more efficient in classifying classification tasks into two or more homogeneous sets. The classification tasks depend on the target variables. Using a decision tree has various advantages: easy to understand, faster, less data cleaning, and non-parametric (doesn't require assumptions about the classifier or spatial distribution). However, the main problem of decision trees is over-fitting, which occurs when a model learns the details and noise of the training dataset as a concept and applies it to new data. The new concepts or practices affect the model's generalization ability negatively, reducing the overall model's performance. It is worth considering that non-parametric algorithms are more flexible to learn and are subject to overfitting (Sullivan (2017), Yiu (2021)).

Concerning the regression task, the DT prediction value at the terminal node will be the mean response value of the subgroup. And in the classification task, the predicted class at the terminal node will be the observation mode. Both tree processes are top-down, and the splitting process on nodes continues until the algorithm meets the user's stopping criteria such as depth of the tree, maximum number of terminal nodes, minimum sample for node split, and other. This leads to model overfitting and is the cause of accuracy redaction.

The random forest model is a popular ensemble-supervised machine learning algorithm that combines many tree-based predictors or classifiers. The logical diagram of the RF model is shown in Fig. 6. An ensemble model is a set of weak predictive models trained independently to transform weak learners into strong or more robust ones. The ensemble prediction value will be the combined prediction values of all week's models, done by boosting or bagging methods. The idea behind this is to trade off between bias and variance error; the more complex the model is, the higher the variance and less bias will be, whereas the less variance is, the higher the bias. Hence, the ensembling method comes into the equation to find a balance point(Sullivan (2017)).

To discuss further, boosting is a sequential learning technique that transforms weak models into strong ones by iteratively improving upon the errors (XGBoost, Gradient Boosting, and AD Boost are good examples). However, in the bagging method, decision trees are created to classify objects based on a sub-training set of the dataset and the attributes; each tree presents its classification prediction called "vote".

4.3 Extreme Learning Machine

Extreme Learning Machine is one of the feedforward Neuron Networks (FFNNs) with a new, faster learning technique for a machine system to imitate human behavior(Aku-

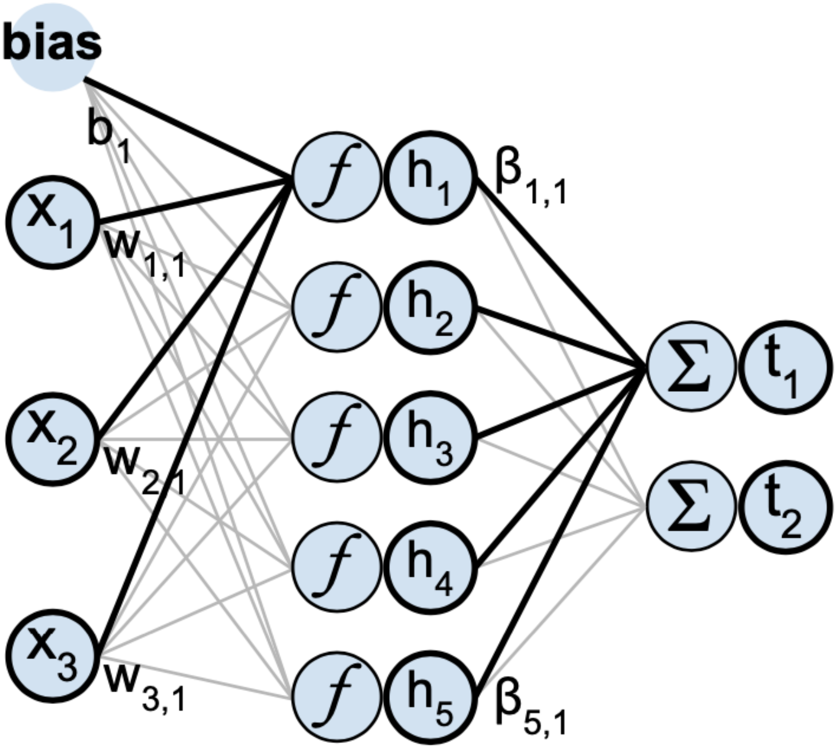


Figure 7. Single Layer Feedforward Network (SLFN) for ELM.

sok et al. (2019b), Leal et al. (2018)). The single-layer feedforward Neuron Networks (SLFNs) with hidden layer, bias function, and adjustable parameters has universal approximator property. The conventional backpropagation (BP) learning technique is time-consuming and prone to overfitting. Technically, BP is a repetitive process of calculating and minimizing loss function based on the weight and bias parameters to optimize the weight coefficient of the relationship between the input and output layers of the hidden block.

In ELM, the parameters of the hidden block (input weights, biases of additive neurons, and others) do not require to be tuned; instead, they are randomly generated independent of the input data. The learning process is feedforward and non-iterative; unlike backpropagation, it's more stable and generalizes the new data with better accuracy. It has been noticed that ELM provides solutions 5 times faster than Multilayer Perceptron (MLP) or 6 times faster than Support Vector Machines (SVM). Fig. 7 illustrates single layer feedforward neuron network for the ELM model, source, and for further reading (Akusok et al. (2015), Deng et al. (2015)).

The diagram in Fig. 7 has three components: the input component X, the hidden component h, and an output component t. The coefficients W, b, and β represent the input weight, the bias (error component), and the output weight, respectively. It is in the hidden layer where data projection (using the weight and bias) and transformation of the projected data is made to generate the output layer's weight (β) and then, finally, on the output layer, a prediction value (t) (Akusok et al. (2015)). Hence, as compared with conventional ML and traditional neural network algorithms, ELM offers significant advantages: less learning time, ease of implementation, good generalization, and takes minimal human intervention.

The mathematical formula of ELM model estimation with L number of hidden layers and with N number of input features is written as follows (Akusok et al. (2015), Burnpiro (2020)):

$$F_L(x) = \sum_{i=1}^{L} \beta_i f_i(x) \tag{1}$$

$$F_L(x) = \sum_{i=1}^{L} \beta_i f(w_i * x_j + b_j)\, where j = 1,..,N \tag{2}$$

Where x is the input vector, b is the bias vector, W is the weight vector between the input and hidden layers, f is the activation function, and β is the weight vector between the hidden and output layers. The transformed data h in the hidden layer is used to generate β, and equation (2) can be shortened as:

$$T = H\beta \tag{3}$$

where

$$H = \begin{bmatrix} (w_1 * x_1 + b_1) & \cdots & (w_l * x_1 + b_l) \\ \vdots & \ddots & \vdots \\ (w_1 * x_1 + b_1) & \cdots & (w_l * x_1 + b_l) \end{bmatrix}_{NL}$$

$$T = \begin{bmatrix} t_1^T & \cdots & t_N^T \end{bmatrix}_{NM}$$

$$\beta = \begin{bmatrix} \beta_1^T & \cdots & \beta_L^T \end{bmatrix}_{LM}$$

M is the number of outputs, H is the hidden layer output matrix, and T is the training data target matrix. The estimated output weight $\hat{\beta}$ using the Moore-Penrose generalized inverse is as follows:

$$\hat{\beta} = TH^{+} \tag{4}$$

Where H^{+} is the Moore-Penrose (MP) generalization inverse of matrix H.

Generally, the learning and prediction process of the ELM model is easy, and the steps are presented as follows.

1. Assigin the input weight W_i and bias b_i randomly, $i = 1 \ldots L$.

2. Calculate hidden layer output H.

3. Calculate output weight $\hat{\beta}$

4. Predict T on new data

To explore the benefits of ELM, both models were implemented to predict soil types as ASS or non-ASS and compare the models' performance based on their classification accuracy.

4.4 Model Selection And Model Training

The next phase after prepossessing is model selection and training using the training dataset, along with parameter tuning. There are several machine learning algorithms to choose from. Generally, the selection process relies on the dimension of the dataset, the required accuracy, the interpretability of the output, the time needed to train, and the linearity of the training dataset. This experiment employed two classification models: Extreme Learning Machine (ELM) and Random Forest (RF). The dataset was partitioned into train and test sets; the train-to-test ratio choice was 3 to 1. The test set is meant to validate the predictive performance of the fitted model.

Technically speaking, a hyperparameter is a high-level attribute: like n_job, n_estimators, max_depth, that a practitioner sets before model training. Besides that, the model learns other characteristics by finding a mathematical relationship between the training dataset (features and target variable).

This research employed a randomized cross-validation search for hyperparameter tuning to get the best cross-validation score for the RF model. The optimization process of parameter tuning is carried out to archive a better accuracy model prediction of soil types, ASS or non-ASS. The parameters tuning phase depends on manually seated parameters space with a randomized cross-validation search of 10 folds. That means 10 randomly categorized subset groups were created from the training dataset, and each of the subsets of the train groups was used to validate the model performance; the rest 9 subsets were used on training the model. Therefore, because of the deployment of the cross-validation search technique, 10 different models were fitted with the corresponding 10 sets of validation estimators. Through this process, the best score's parameters among the 10 fitted model parameters are selected and used for building RF model prediction of soil types, ASS or non-ASS.

In the case of ELM, there is no need to carry out the time-consuming hyper-parameter tuning task. That is one of the benefits of using ELM solutions for machine learning prediction tasks. Scikit-ELM toolbox was used because of its flexibility and usability; the reader is directed to the canonical papers for more detail(Akusok et al. (2015, 2019a)).

5 PROJECT FRAMEWORK

The project framework illustrates the whole end-to-end machine Learning process of model building, from data ingestion to model evaluation in predicting soil types. As Fig. 8 below demonstrates, the model development process that includes many complicated data analysis and image processing tasks. The tasks are dataset collection (point observations and environmental covariates layers), model selection, model training, parameter tuning, and assessing the performance of the choice model.

5.1 Data Preparation

The sample point dataset is prepossessed by creating GeoPandas dataframe with a new variable of point geometry. Since Finland is located near the north pole, the point dataset's longitude and latitude need to be set to the regional standard coordinate reference system (CRS) "WGS84" to avoid image distortion that occurs near the pole. The next step will be extracting pixel coordinates (x,y) and tile coordinates (z,x,y) using the geographical location coordinates of the sample point(longitude and latitude) and zoom level z. Tile image (z,x,y) is a 256 X 256 pixel-sizes of multiple neighboring pixel points (x,y) for a given zoom level z.

5.2 Data Encoding

The research project is a supervised binary classification of ASS and non-ASS. The target feature "class" is a categorical feature that needs to be converted to a numerical feature of 0 and 1 for soil type ASS and non-ASS, respectively.

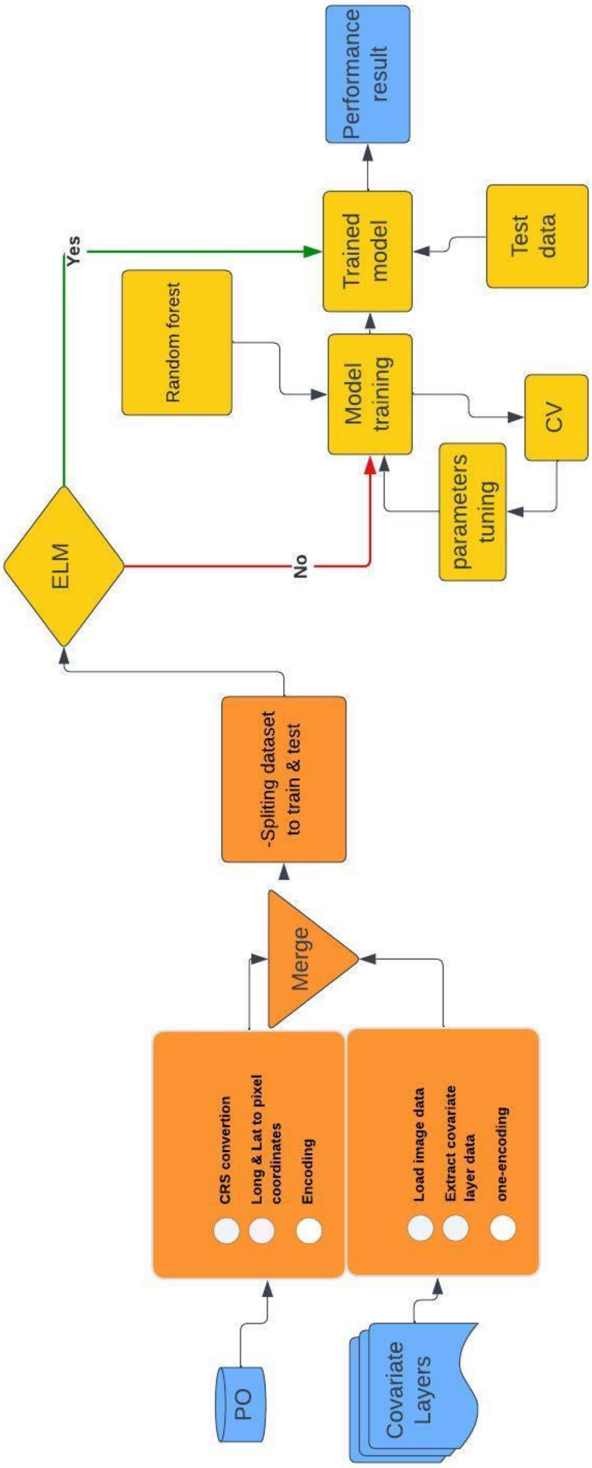


Figure 8. Model development framework for acid sulfate soil prediction

The prepossessing task of covariate layers is complex, and it starts by loading tiles images of the 13 environmental covariate layers. A Python function was developed to load the image tiles and extract the covariate layer's data pixel for a given sample point location and zoom level. The two covariate layers, corine-land-cover and bedrock are categorical variables; hence, the one-hot encoding technique was used to represent the categorical values of the variables. The encoding method helps to avoid the ordinal relationship of integral values between values used to distinguish the categorical attributes of the given feature. Other prepossessing works include merging the two datasets (points and covariate layers datasets), splitting the database into train and validation sets, and scaling the training dataset using the Scikit-learn RobustScaler module.

5.3 Feature Selection

Feature selection is one of the important data processing techniques for selecting features that contribute most to model building. In most cases, incorporating irrelevant or less significant features in model training has a negative impact on the generalization role of the model on unseen (Estévez et al. (2023)). Removing the irrelevant or less informative features reduces over-fitting, improves performance, and decreases training time. Fig. 9 illustrates features importance of the Random Forest model for the entire 103 features of the study database.

Generally, more features in the model mean more complexity and more time to train a model. It is vital to reduce the size of the features to a level that would not harm the model's performance. This experiment

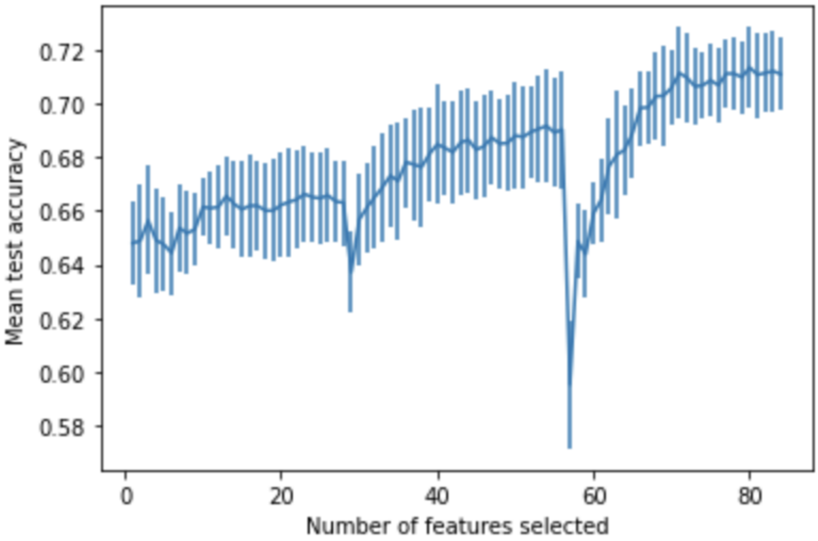


Figure 9. Accuracy of Random Forest model with respect to the number of features

employed a Scikit-learn tool called RFECV (Recursive Feature Elimination Cross-Validation), with 10 folds cross-validation features selection method. RFECV works on a subset of all possible space of features, recursively training a model and pruning the less significant one on all possible subsets until the optimal number of features is reached. Deploying the RFECV method using a random forest model on the study database reduces the feature number to 22. Fig. 10 shows the selected features and their importance.

It is apparent that elevation, pixel x, pixel y, and aem_imaginary features are the most significant features and account for about 44% of the RF model's predictive power. Features aspect, NDVI, TRI, TWI, and aem_real in aggregate accounts for about 23%, and features hillshade, slope, TPI, and aem_apparent contribute about 15% of the predictive power of the RF model. Therefore, out of 103 features, 12 features contribute about 82% of the total predictive power of the RF model.

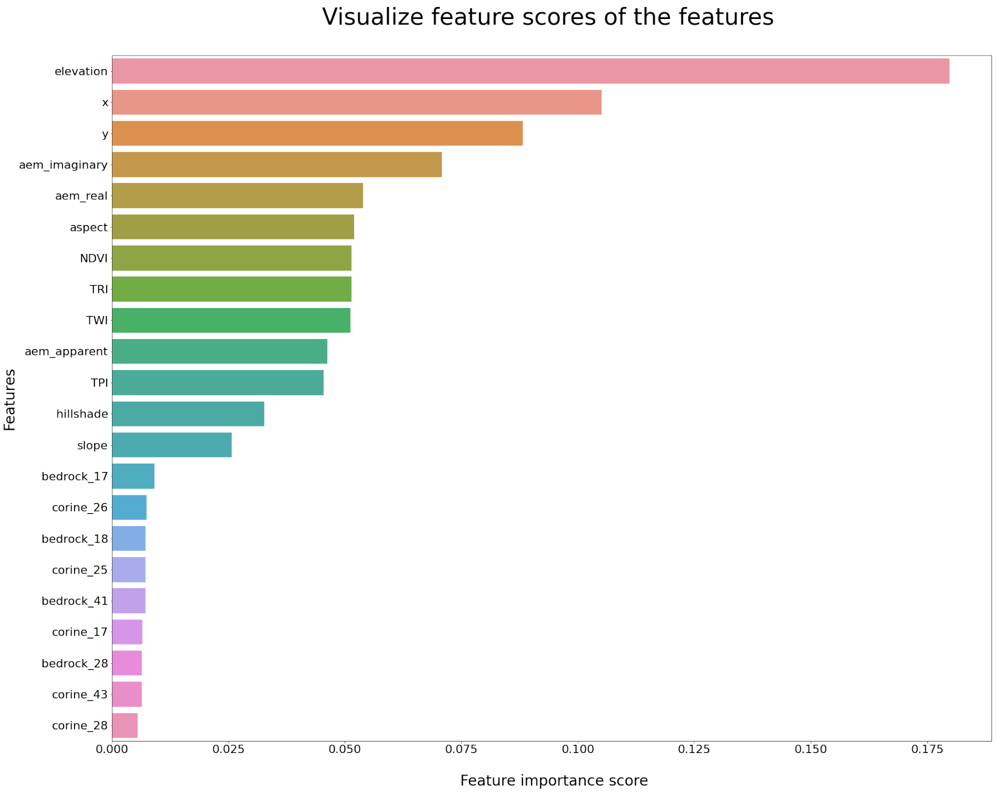


Figure 10. Random Forest feature importance for the most informative features

6 EXPERIMENTAL RESULTS

The thesis project explores the ELM's classification of ASS using the environmental covariates variables based on the sample point observations taken from the west coast area of Finland. The following section presents the results of the project experiments obtained using machine learning models, namely Random Forest(RF) and Extreme Learning Machine(ELM). Before presenting the results, it is essential to discuss the metrics used for evaluating the results of the experiments and the methods used.

6.1 Evaluation Metrics

The overall objective of building a predictive machine learning model is to deliver a high accuracy score for unseen data. They are measuring how robust the model prediction is, and explaining the performance before deployment is essential. Several evaluation metrics exist, and their selection depends on the model types and the implementation plan. Some of the evaluation matrices are discussed as follows (Estévez Nuño (2020)):

- Confusion Matrix: is N X N matrix, where N is the number of the predicted classes, describes the complete performance of the model. Table2 shows a confusing matrix where TP is true positive, FN is false negative, FP is false positive, and TN is true negative.

  Besides that, the confusion matrix helps to drive important measures: precision-recall, accuracy, and AUC_ROC curve.

Table 2. Confusing Matrix

		Predictive Values	
Actual Values	Positive	TP	FN
	Negative	FP	TN

- precision or Sensitivity: the proportion of positive cases correctly identified, i.e.

$$\frac{TP}{(TP+FP)}$$

- Recall or specificity (negative predictive value): is the proportion of correctly identified negative cases, i.e.,

$$\frac{TP}{(TP+FN)}$$

- Classification Accuracy: the proportion of correct predictions to the total number of input samples, i.e.,

$$\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

- F1-score: represents a balanced combination of precision and recall, taking the harmonic mean of these two metrics. An F1 score of 1 is the optimal value and 0 is the lowest value, i.e., (Pedregosa et al. (2011))

$$2 * (\frac{precision * Recall}{precision + Recall})$$

- The AUC_ROC (Area Under Curve _ Receiver Operating Characteristics) curve: represents the true positive rate (Sensitivity) as a function of the false positive rate (1-specificity). The AUC value is between 0 and 1; if the value is above threshold 0.5, the model can identify the classes very well.

6.2 Results And Discussion

The research experiment used RF and ELM models for class prediction of soil as acid sulfate soil (ASS) or normal soil (non-ASS). The evaluation statistics are described as follows. In the case of RF, Fig. 11 compares mean test score versus tree size (param_n_estimators) for tree depth (param_depth) of 10 and 20. The model offers a higher mean_score for 20 param_depth than 10 for a given number of trees (param_n_estimator).Fig. 12 shows that the higher the tree depth, the better the performance is, but with a more extended period to process.

As discussed in the previous sections, randomized grid searches for hyper-parameter tuning and feature selection were deployed on the RF model to get the best performance. The best results are shown in Fig. 13, and this figure presents a confusion matrix and related metrics called classification reports. The confusion matrix presents the RF model correctly classifying 322 sample points as Normal soils (True Positive) and 705 as Acid Sulfate soil (True Negative). However, the model mispredicted 171 sample soils as

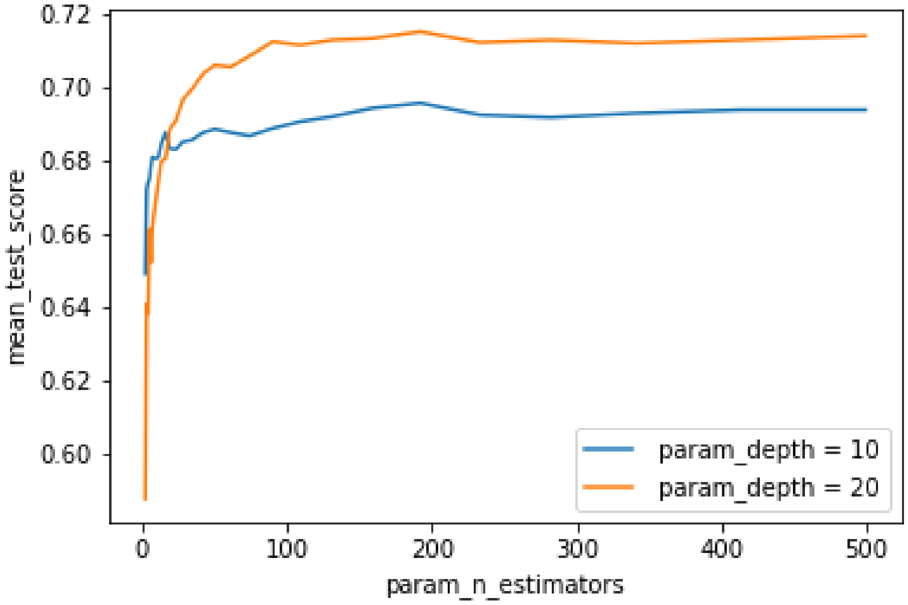


Figure 11. Random Forest test score line chart with 10 and 20 tree depths

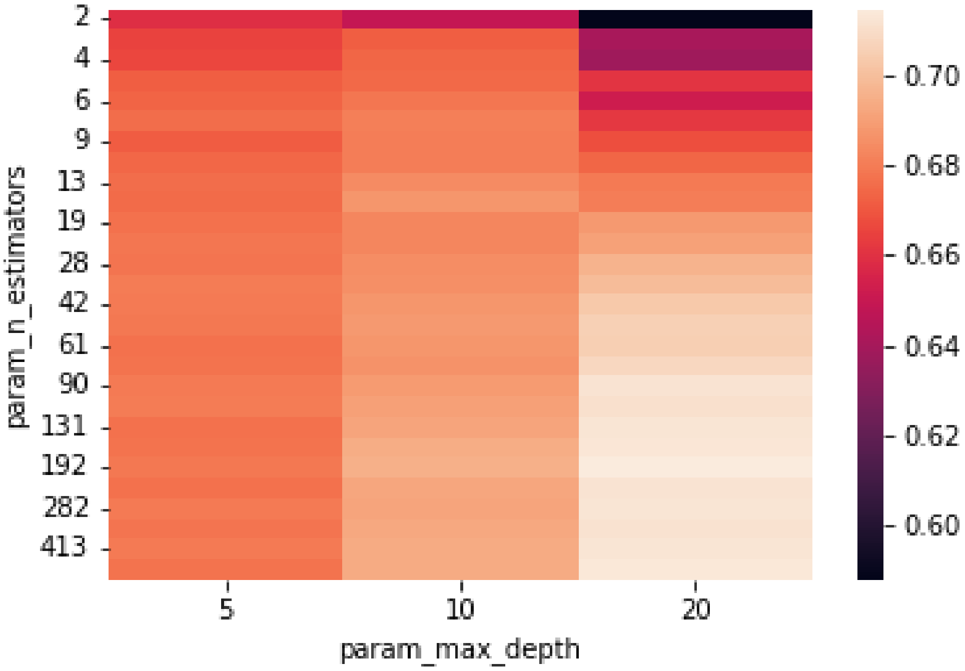


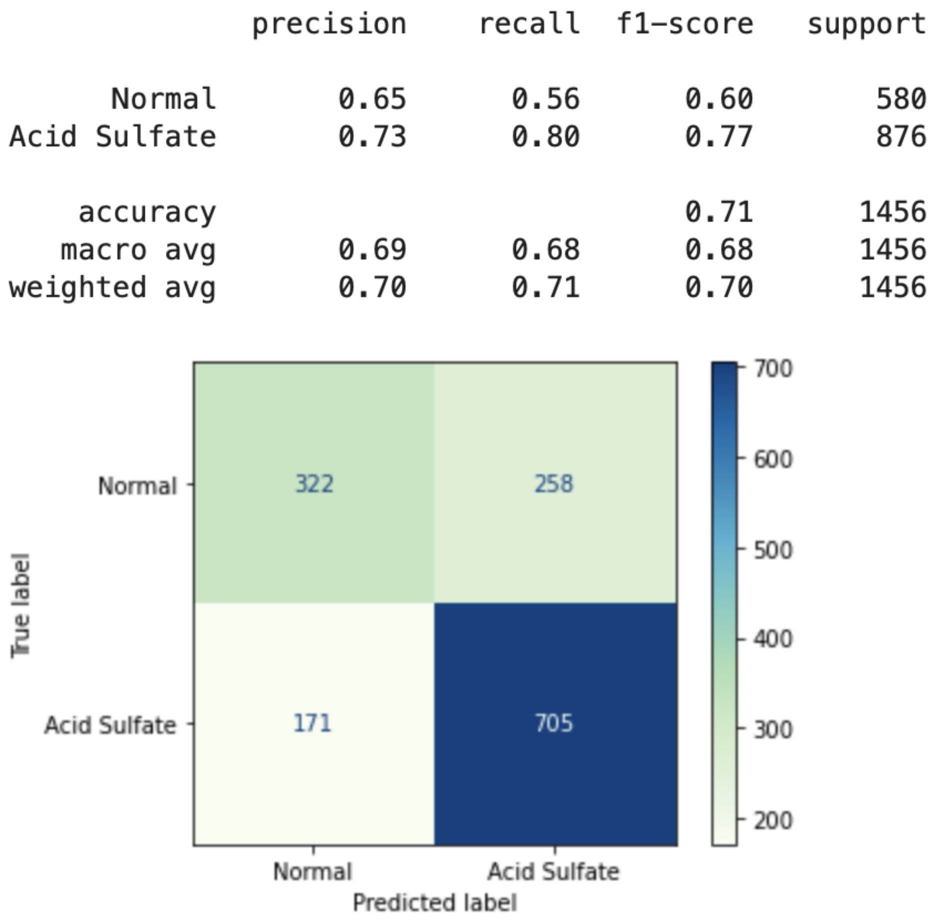
Figure 12. Heatmap of Random Forest test score, decision trees versus tree depth

Normal and 258 as Acid Sulfate soil. Both classes' correct predicted soil types are greater than the incorrectly predicted classes. Therefore, the RF model built is capable of classifying unseen location soil with significant accuracy.

The classification report presented at the top of the confusion matrix shows the overall accuracy of the RF model is 71%, i.e., 71% of the prediction is correct. The precision for Acid Sulfate and Normal soil types are 73% and 65%, respectively, i.e., 73% and 65% of the respective soil types classified as such are accurate. However, the recalls are 80% and 56% for Acid Sulfate and Normal soil types, respectively, i.e., 80% and 56% of all soil types of the respective soils are classified correctly. The prediction for each class is presented by F1-score 77% for Acid Sulfate and 60% for the Normal, so the model classifies Acid Sulfate soils with better accuracy than the Normal.

Fig. 14 depicts the ROC curve representing the performance of the RF model. In this illustration, the green curve surpasses the blue one(non-discrimination line), signifying the model's effective classification. The ROC curve's area under it is 0.78, which is not a perfect case where ROC equals 1. The model prediction rank of 0.78 reflects the RF model prediction is significant.

ELM model Implementation of acid soil classification using the Scikit-elm library is very easy (Akusok et al. (2019a)). As shown in the project framework section of Fig. 8, in the ELM model, unlike conventional learning algorithms, the parameters of the hidden layers are randomly established and



```
              precision    recall  f1-score   support

      Normal       0.65      0.56      0.60       580
Acid Sulfate       0.73      0.80      0.77       876

    accuracy                           0.71      1456
   macro avg       0.69      0.68      0.68      1456
weighted avg       0.70      0.71      0.70      1456
```

Figure 13. Classification report and confusion matrix for Random Forest model

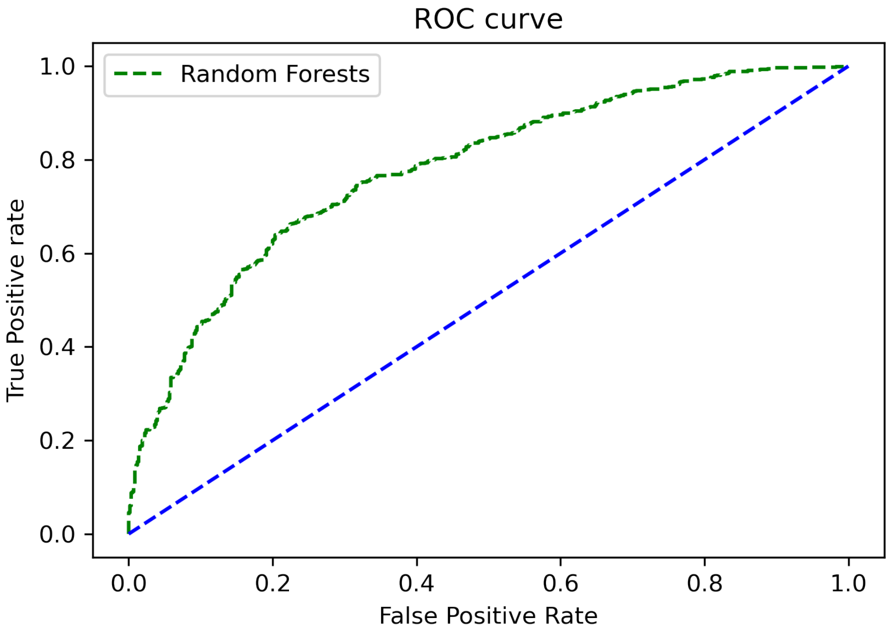
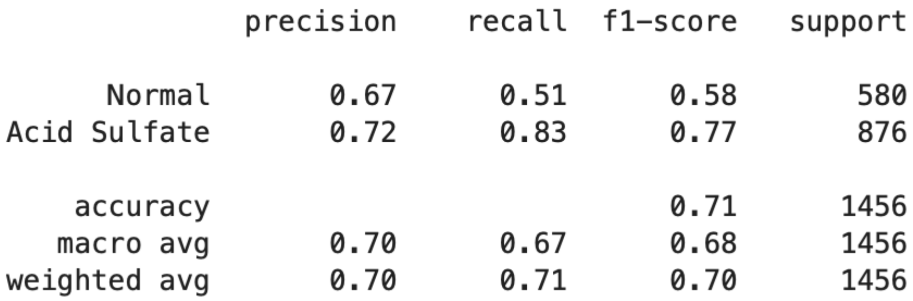


Figure 14. ROC curve for Random Forest

don't need to tune, and it means the training of the hidden nodes is accomplished before the input is acquired (Deng et al. (2015)). Practically, the model is fitted with features and target variables, then the class prediction of soil type ASS or Normal is delivered. The statistics results are shown in Fig. 15; the figure presents the confusion matrix and the classification report of the ELM model. The overall model accuracy is 71%, which is the same percentage as the RF model prediction. The precisions are 72% and 67%, and the recalls are 83% and 51% for Acid Sulfate and Normal soils, respectively. F1-scores are 77% and 58% for Acid Sulfate and Normal soil, respectively, so as the RF model, ELM classifies the Acid Sulfate soil type better than the Normal soil type.



```
              precision    recall  f1-score   support

      Normal       0.67      0.51      0.58       580
Acid Sulfate       0.72      0.83      0.77       876

    accuracy                           0.71      1456
   macro avg       0.70      0.67      0.68      1456
weighted avg       0.70      0.71      0.70      1456
```

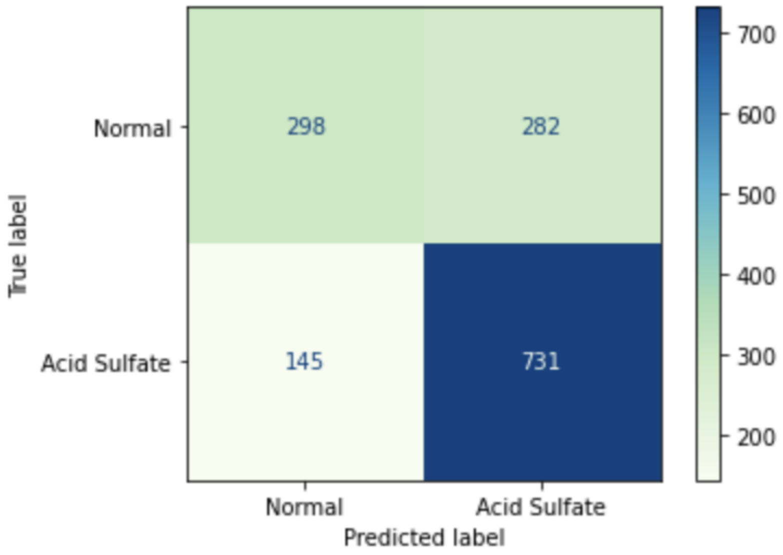


Figure 15. Classification report and confusion matrix for ELM model

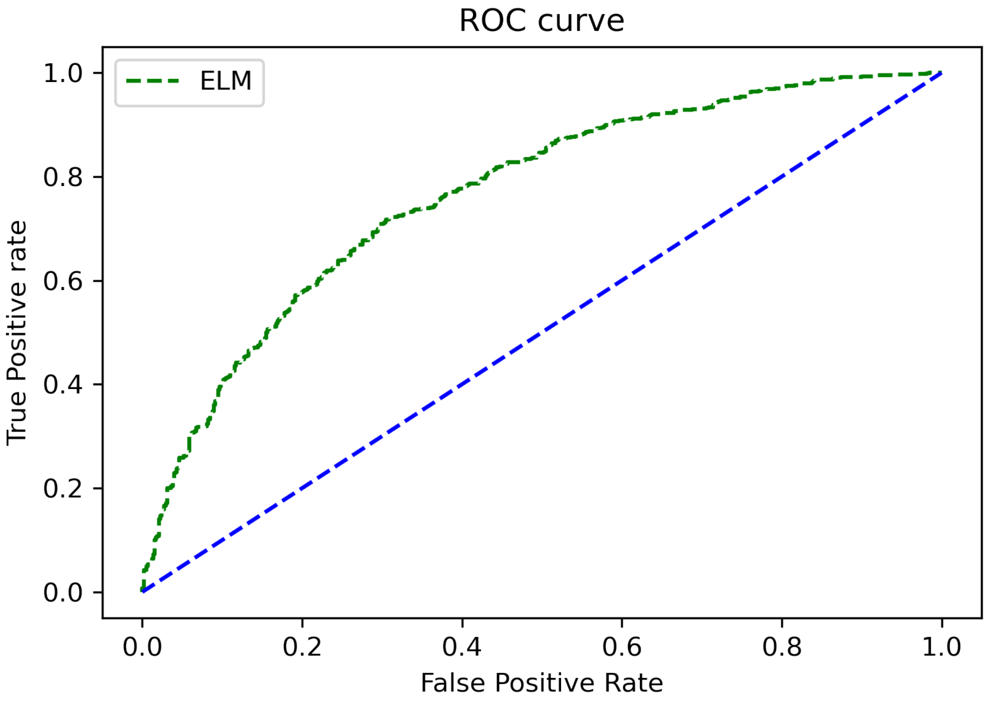


Figure 16. AUC-ROC for Extreme Machine Learning

The ELM ROC curve in Fig. 16 is above the non-discrimination line, which means the model works well in classifying the soil types. The area under the curve ROC is 0.76, it is not a perfect result, but the rank explains that the ELM model prediction is significant.

Comparing the two models based on the statistical results presented above, the precision values of both models indicate that both models performed almost equally the same in predicting ASS (Acide Sulfate). Still, the ELM model performs better predicting non-ASS(Normal) type. About recalls: concerning prediction on its own class, ELM performed better in predicting ASS, and RF performed better in predicting non-ASS type. However, the overall performance of both models is the same. On the other hand, comparing the processing time and complexity in the implementation process, the ELM model is very fast and user-friendly.

7 CONCLUSIONS

This project aims to categorize soil as either ASS or non-ASS based on environmental covariates represented by layers of tiles, utilizing ELM algorithm. Besides that, the research focuses on identifying the most relevant features and exploring the performance of ELM and RF models. To conduct this experiment, sample point observations dataset and environmental covariates layers of images were employed.The experiment incorporated 13 environmental covariates layers and two locational variables, longitude and latitude. During the prepossessing phase, one-hot encoding techniques were utilized, and additional features were generated, thus increasing the overall feature count to 103.

After data processing, RF and ELM classification models were implemented to predict soil type. The implementation of the Scikit-learn ELM model doesn't require hyperparameter tuning, whereas RF requires a randomized grid search hyperparameter tuning to get the best performance. Additionally, feature selection techniques were used to enhance the RF model's performance by removing the less significant and irrelevant features in predicting soil class. As a result, 22 features were chosen based on their level of influence on the RF model's predictive ability. Both models work well in classifying soils as either ASS or non-ASS, but neither is perfect. Notably, the ELM model stands out for its swiftness and user-friendliness. One potential research avenue in the future will be optimizing the model performance by incorporating additional environmental covariates.

REFERENCES

Akusok, A; Björk, K; Miche, Y & Lendasse, A. 2015, High-performance extreme learning machines: a complete toolbox for big data applications, *IEEE Access*, vol. 3, , pp. 1011–1025.

Akusok, A; Björk, K; Estévez, V & Boman, A. 2023, Randomized Model Structure Selection Approach for Extreme Learning Machine Applied to Acid Sulfate Soil Detection, pp. 32–40.

Akusok, A; Leal, L Espinosa; Björk, K & Lendasse, A. 2019a, High-performance ELM for memory constrained edge computing devices with metal performance shaders, In: *International Conference on Extreme Learning Machine*, Springer, pp. 79–88.

Akusok, A; Leal, L Espinosa; Björk, K & Lendasse, A. 2019b, Scikit-ELM: an extreme learning machine toolbox for dynamic and scalable learning, In: *International Conference on Extreme Learning Machine*, Springer, pp. 69–78.

Andreas C.Müller, Sarah Guido. 2016, *Introduction to Machine Learning with Python*, O'Reilly Media.

Andriesse, W. & van Mensvoort, M.E.F. 2002, *Acid sulfate soils, distribution and extent*, Marcel Dekker, p. 6.

Baltensweiler, A; Walthert, L; Hanewinkel, M; Zimmermann, S & Nussbaum, M. 2021, Machine learning based soil maps for a wide range of soil properties for the forested area of Switzerland, *Geoderma Regional*, vol. 27, , p. e00437. Available: https://www.sciencedirect.com/science/article/pii/S2352009421000821.

Bernard M, Matt W. 2019, *Artificial Intelligence in Practice (1st ed.)*, Wiley. Available: https://www.perlego.com/book/991892/artificial-intelligence-in-practice-how-50-successful-companies-used-ai-and-machine-learning-to-solve-problems-pdf(Originalworkpublished2019).

Beucher, A.; Siemssen, R.; Fröjdö, S.; Österholm, P.; Martinkauppi, A. & Edén, P. 2015, Artificial neural network for mapping and characterization of acid sulfate soils: Application to Sirppujoki River catchment, southwestern Finland, *Geoderma*, vol. 247-248, , pp. 38–50. Available: https://www.sciencedirect.com/science/article/pii/S0016706115000464.

Brevik, E; Slaughter, L; Singh, B; Steffan, J; Collier, D; Barnhart, P & Pereira, P. 2020, Soil and Human Health: Current Status and Future Needs, *Air, Soil and Water Research*, vol. 13, , p. 117862212093444.

Burnpiro, K. 2020, *Introduction to extreme learning machines*. Available: https://towardsdatascience.com/introduction-to-extreme-learning-machines-c020020ff82b.

Deenik, K. 2021, *Topographic wetness index (TWI) – part One*. Available: https://blogs.ubc.ca/tdeenik/2021/03/08/topographic-wetness-index-twi/.

Deng, C; Huang, G; Xu, J & Tang, J. 2015, Extreme learning machines: new trends and applications, *Science China Information Sciences*, vol. 58, , pp. 1–16.

Domingos, P. 2012, A Few Useful Things to Know about Machine Learning, *Commun. ACM*, vol. 55, no. 10, p. 78–87.

Eash, Neal S; Sauer, Thomas J; O'Dell, D & Odoi, E. 2015, *Soil science simplified*, John Wiley & Sons.

Eden, P; Weppling, K & Jokela, S. 1999, *Natural and land-use induced load of acidity, metals, humus and suspended matter in Lestijoki, a river in western Finland.* Available: https://www.borenv.net/BER/archive/pdfs/ber4/ber4-031-043.pdf.

Epie, Kenedy E.; V, Seija; S, Arja; S, Asko & Stoddard, Frederick L. 2014, The effects of a permanently elevated water table in an acid sulphate soil on reed canary grass for combustion, *Plant and Soil*, vol. 375, no. 1-2, pp. 149–158.

Espinosa-Leal, L; Chapman, A & Westerlund, M. 2020, Autonomous industrial management via reinforcement learning, *Journal of intelligent & Fuzzy systems*, vol. 39, no. 6, pp. 8427–8439.

Estévez, V; Beucher, A; Mattbck, S; Boman, A; Auri, J; Björk, K & Österholm, P. 2022, Machine learning techniques for acid sulfate soil mapping in southeastern Finland, *Geoderma*, vol. 406, , p. 115446. Available: https://www.sciencedirect.com/science/article/pii/S0016706121005267.

Estévez, V; Mattbäck, S; Boman, A; Beucher, A; Björk, K-M & Österholm, P. 2023, Improving prediction accuracy for acid sulfate soil mapping by means of variable selection., *Front. Environ. Sci.11:1213069.* Available: https://www.frontiersin.org/articles/10.3389/fenvs.2023.1213069/full.

Estévez Nuño, V. 2020, Machine Learning Methods for Classification of Acid Sulfate Soils in Virolahti, *Master's thesis.* Available: https://urn.fi/URN:NBN:fi:amk-2020052915446.

Evans, Jeffrey S. *Terrain ruggedness index.* Available: https://search.r-project.org/CRAN/refmans/spatialEco/html/tri.html.

F, Rasmus; Åström, Mats & Vuori, K M. 2008, Environmental risks of metals mobilised from acid sulphate soils in Finland: a literature review, *Boreal Environment Research*, vol. 13, , pp. 444–456.

Forrest, M. 2023a, *A brief history of web maps.* Available: https://forrest.nyc/a-brief-history-of-web-maps/.

Forrest, M. 2023b, *Map tiles: Everything you need to know.* Available: https://carto.com/blog/map-tiles-guide.

GISGeography. 2022, *What is NDVI (normalized difference vegetation index)?* Available: https://gisgeography.com/ndvi-normalized-difference-vegetation-index/.

Huang, P.M.; Li, Y. & Sumner, M.E. 2011, *Handbook of Soil Sciences: Resource Management and Environmental Impacts, Second Edition*, Handbook of Soil Science, Taylor & Francis.

Jaakko, A; M, Stefan; B, Anton; L, Pauliina; R, Jukka & H, Hannu. 2022, *From a general survey to risk management – acid sulfate soils are Finland's most persistent environmental problem, but research can mitigate the harms they cause*. Available: https://www.gtk.fi/en/current/from-a-general-survey-to-risk-management-acid-sulfate-soils-are-finlands-most-persistent-environmental-problem-but-research-can-mitigate-the-harms-they-cause.

Joukainen, S & Yli-Halla, M. 2003, Environmental impacts and acid loads from deep sulfidic layers of two well-drained acid sulfate soils in western Finland, *Agriculture, Ecosystems & Environment*. Available: https://www.sciencedirect.com/science/article/pii/S0167880902000944.

L, Rattan; B, Johan; B, Eric; D, Lorna; Field, Damien J.; G, Bruno; H, Ryusuke; Hartemink, Alfred E.; K, Takashi; L, Bruce; M, Curtis; M, Cristine; N, Georges M.; N, Stefan; P, Xicai; P, Remigio; B, Laura; S, Taru; S, Bal R.; S, Heide; Y, Junta & Z, Jiabao. 2021, Soils and sustainable development goals of the United Nations: An International Union of Soil Sciences perspective, *Geoderma Regional*, vol. 25, , p. e00398. Available: https://www.sciencedirect.com/science/article/pii/S2352009421000432.

Lal, Rattan. 2017, *Encyclopedia of SoilScience*, Boca Raton.

Leal, L Espinosa; Björk, K; Lendasse, A & Akusok, A. 2018, A web page classifier library based on random image content analysis using deep learning, In: *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, pp. 13–16.

Li, Y; Yang, R & Guo, P. 2020, Spark-based parallel OS-ELM algorithm application for short-term load forecasting for massive user data, *Electric Power Components and Systems*, vol. 48, no. 6-7, pp. 603–614.

MapTiler. 2023, *Tiles à la google maps: Coordinates, tile bounds, and projection*. Available: https://www.maptiler.com/google-maps-coordinates-tile-bounds-projection.

McBratney, A.B; Mendonça Santos, M.L & Minasny, B. 2003, On digital soil mapping, *Geoderma*, vol. 117, no. 1, pp. 3–52. Available: https://www.sciencedirect.com/science/article/pii/S0016706103002234.

Minasny, Budiman & McBratney, Alex.B. 2016, Digital soil mapping: A brief history and some lessons, *Geoderma*, vol. 264, , pp. 301–311, soil mapping, classification, and modelling: history and future directions. Available: https://www.sciencedirect.com/science/article/pii/S0016706115300276.

Ministry of Agriculture and Forestry Ministry of the Environment. 2011, Guidelines for mitigating the adverse effects of acid sulfate soils in Finland until 2020. Available: http://urn.fi/URN:ISBN:978-952-453-741-4.

Palko, J. 1986, Mineral Element Content of Timothy (Phleum pratense L.) in an Acid Sulphate Soil Area of Tupos Village, Northern Finland, *Acta Agriculturae Scandinavica*, vol. 36, no. 4, pp. 399–409. Available: https://doi.org/10.1080/00015128609439897.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. & Duchesnay, E. 2011, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, vol. 12, , pp. 2825–2830.

Riley, S; Degloria, S & Elliot, S.D. 1999, A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity, *Internation Journal of Science*, vol. 5, , pp. 23–27.

Sarangi, S; Mainuddin, M & Maji, B. 2022, Problems, Management, and Prospects of Acid Sulphate Soils in the Ganges Delta, *Soil Systems*, vol. 6, no. 4, p. 95.

Sharma, A. 2023, *4 simple ways to split a decision tree in Machine Learning (updated 2023)*. Available: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/.

Sullivan, W. 2017, *Machine learning beginners guide algorithms*, North Charleston, SC: Createspace Independent Publishing Platform.

SYKE. *Corine maanpeite 2018 - Syke*. Available: https://ckan.ymparisto.fi/dataset/corine-maanpeite-2018.

Yiu, T. 2021, *Understanding random forest*. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

Yli-Halla, M. 2022, Acid sulfate soils: A challenge for environmental sustainability, *Annales Academiae Scientiarum Fennicae. Geologica-Geographica*, vol. 1, no. 1, pp. 124–141. Available: http://hdl.handle.net/10138/350980.

Yli-Halla, M; Puustinen, M & Koskiaho, J. 1999, Area of cultivated acid sulfate soils in Finland, *Soil Use and management*, pp. 62–67. Available: https://doi.org/10.1111/j.1475-2743.1999.tb00065.x.

Čučković, Z. 2019, *Terrain Position index for QGIS*. Available: https://landscapearchaeology.org/2019/tpi/.