since 1961

Baltica

BALTICA Volume 37 Number 2 December 2024: 137–150
https://doi.org/10.5200/baltica.2024.2.5

Hydraulic conductivity determination of Lithuanian soils using machine learning

Vytautas Samalavičius*, Eveliina Kukka-Maaria Vanhala, Ieva Lekstutytė, Sonata Gadeikienė, Saulius Gadeikis, Gintaras Žaržojus

Samalavičius, V., Vanhala, E.K.-M., Lekstutytė, I., Gadeikienė, S., Gadeikis, S., Žaržojus, G. 2024. Hydraulic conductivity determination of Lithuanian soils using machine learning. *Baltica 37 (2)*, 137–150. Vilnius. ISSN 1648-858X.
Manuscript submitted 10 August 2024 / Accepted 27 November 2024 / Available online 9 December 2024

© Baltica 2024

Abstract. Hydraulic conductivity (*k*) is a crucial parameter in hydrogeology and engineering geology, describing the rate at which water filters through porous media, i.e., soil. It can be determined directly through tests on soil samples in situ, or it can be calculated from other soil parameters using various equations and models. This study aims to compare the results of six machine learning (ML) models with those of four empirical formulas and to identify the soil parameters required for the optimal ML performance. A dataset consisting of 282 unique entries of Lithuanian soils was compiled from laboratory testing reports. Twelve features, including grain sizes and particle diameters, were used to create 4095 combinations of inputs for each ML algorithm. Prediction results were evaluated using the determination coefficient (R^2) and the mean absolute error (MAE). The ML models provided more accurate predictions (R^2 0.36–0.46, MAE 2.31–2.81 m/d) compared to the empirical formulas (R^2 0.10–0.33, MAE 3.05–6.54 m/d). However, some ML models showed signs of overfitting. The study also revealed that each ML algorithm performs best with a customized combination of input parameters, ranging from 4 to 8, whereas the empirical formulas used in this study utilize only 1–2 parameters.

Keywords: *permeability; grain size; particle diameter; artificial intelligence; groundwater*

Vytautas Samalavičius (vytautas.samalavicius@chgf.vu.lt), https://orcid.org/0000-0001-5791-7483*
Eveliina Kukka-Maaria Vanhala (eveliina.vanhala@hotmail.com), https://orcid.org/0009-0007-3163-4216
Ieva Lekstutytė (ieva.lekstutyte@chgf.vu.lt), https://orcid.org/0000-0003-0104-0605
Gintaras Žaržojus (gintaras.zarzojus@gf.vu.lt), https://orcid.org/0000-0003-2079-4680
Sonata Gadeikienė (sonata.gadeikyte@gf.vu.lt), https://orcid.org/0000-0001-6504-5769
Saulius Gadeikis (saulius.gadeikis@gf.vu.lt), https://orcid.org/0000-0001-5054-440X
Department of Hydrogeology and Engineering Geology, Vilnius University, Vilnius, Lithuania
**Corresponding author*

INTRODUCTION

Hydraulic conductivity (*k*) is a property of soil[1] that allows fluid (e.g. water) to flow through the interconnected pores between the soil particles (Alyamani, Şen 1993). A higher *k* value makes the soil more permeable to water. Hydraulic conductivity is an essential parameter in groundwater flow studies and engineering geology. Its applications cover many fields, such as aquifer properties and modelling (Juodkazis *et al.* 2012; Skridlaitė *et al.* 2015), geotechnical design, contaminant migration, and waste disposal (Jang *et al.* 2011).

There are a few possible ways of *k* determination. *In situ* tests are so far the most accurate because they deal with undisturbed soil samples (Dobkevičius 2002; Ibrahim, Aliyu 2016; Skridlaitė *et al.* 2015). Aquifer pump tests give a good estimation of water permeability (Juodkazis *et al.* 2012). However, it is an approximate value of the entire aquifer in question, which may be made from various soils and rocks

1 In engineering geology, the terms "sediments" and "soils" are used interchangeably. Both refer to unconsolidated Earth materials comprising particles like clay, silt, sand, and gravel. What geologists refer to as "Quaternary sediments" are known as "soils" in engineering geological terminology.

with varying k. *In situ* tests are usually expensive and require a substantial workforce and preparation.

Laboratory testing is a common method of k determination. The samples are collected on-site via drilling, digging, or other types of extraction and transported to the laboratory. The most common methods are falling and constant head types (ISO 17892-11 2019; Klizas 2003). In most cases, the samples of loose or coarse soils are disturbed while sampling, which greatly affects the results of k determination due to the loss of their natural internal structure.

A group of empirical methods have been developed to theoretically estimate the hydraulic conductivity of soils. Empirical formulas (EFs) encompass other soil parameters (e.g. porosity, grain size, soil texture and bulk density) to calculate k values.

Numerous variants of EFs have been created to estimate hydraulic conductivity. The theory behind the equations states that particle size distribution affects the effective porosity of the soil, the space where water can freely flow. Soil with a higher content of fine particles (clay, silt) is more compacted and results in a smaller pore size, which restricts water permeability; therefore, grain size distribution is a good proxy for k estimation.

Various authors successfully established their EFs based on particle diameters (d_{xx}), e.g., Hazen, Slichter, and USBR (Hazen 1892; Klizas 2003; Odong 2008; Urumović *et al.* 2020). Particle diameters are obtained from the cumulative grain size distribution curve, where *xx* stands for the percentage, and d is the diameter of a particle at the presented percentage (e. g. d_{10}, d_{50}). The EFs are sometimes restricted to a certain soil type and can only be applied if grain size distribution matches the conditions. For example, the Hazen formula can be applied to soils in which grain size $d_{10} > 0.1$ mm and a uniformity coefficient of $d_{60}/d_{10} \leq 5$ (Klizas 2003). There are also modifications of the Hazen formula for the correction of empirical coefficient C based on a d_{60}/d_{10} of soil. Default C is 100, but the range found in textbooks varies from 1–1000 (Carrier 2003).

This observation reveals a problem with empirical calculations. It is difficult to find a proper equation for universal use. Each EF should be adjusted to a specific soil type, overall grain size distribution, content of the soil minerals, particle shape, etc. Oversimplification of the EF results in the loss of accuracy.

Advances in modern computational techniques may be a solution to the problem. Deep learning (DL) and machine learning (ML) algorithms create a prediction model based on multiple input parameters (features) to achieve the desired output (target). It is an alternative method to the EF for hydraulic conductivity estimation. ML algorithms have an advantage because they can deal with relationships between soil properties and hydraulic conductivity that are complex, non-linear, and involve multiple parameters by means of parametric and non-parametric approaches and neural networks (Li *et al.* 2022).

Successful studies of ML application for k estimation were conducted by many authors covering many soil types and different approaches for ML, data preparation, and models‘ tuning. The results also range from a low correlation between predicted and actual k values ($R^2 < 0.3$) to an almost perfect fit where $R^2 > 0.95$ (Araya, Ghezzehei 2019; Bhaurao More *et al.* 2022; Li *et al.* 2022; Rehman *et al.* 2022; Veloso *et al.* 2022).

Results may depend on the overall dataset size used for modelling, e.g., 29 entries in Li *et al.* (2022) to more than 18,000 (Araya, Ghezzehei 2019). The selection of input features also varies significantly and may include grain size distribution, d_{xx}, soil density, porosity, etc. Various studies focus on research from specific regions’ or countries‘ soil (Veloso *et al.* 2022). Among others, these circumstances lead to the diversity of prediction accuracy and the necessity to create new models for different regions with the specific soil types, accounting for their origin, chemical-mineralogical content, age, etc.

In Lithuania, hydraulic conductivity was thoroughly studied in glacial moraine loam samples (Klizas *et al.* 2015). Soil properties, including k of important areas, were investigated, i.e. clays of radioactive waste repository of the Ignalina nuclear power plant (Klizas 2014) and karst region in north Lithuania (Klizas, Šečkus 2007). The effect of lime introduction on clay soil’s hydraulic conductivity showed that the chemical content of the soil also affects its permeability (Klimašauskas *et al.* 2020).

An extensive study of Lithuanian soil filtration properties by Dobkevičius (2002) included correlations of field and laboratory tests. Research of the total dataset of 48 samples with the k value determined in the laboratory yielded EFs with the R^2 up to 0.96 (Dobkevičius 2002). However, in this study, several EFs were developed for a particular range of grain size distribution using less than a few dozen samples for each, making EFs specifically tailored for a certain type of soil and limiting their applicability.

Extended research conducted by Vanhala (2024) of soil properties application for hydraulic conductivity estimation using the EF and ML by using a dataset (246 unique entries) of diverse types of soils showed a moderate correlation $R^2 < 0.4$ for the EF and $R^2 \sim 0.5$ for ML. The database included a wide range of soils from silt to coarse gravelly sand (Vanhala 2024).

This study aimed to estimate hydraulic conductivity based on grain size distribution and particle size to compare the effectiveness of the EF and ML techniques on Lithuanian soils. A database of samples was developed from the data of laboratory tests conducted in the Department of Hydrogeology and Engineering Geology at Vilnius University. The feature combination approach was applied to select best-fitting dependant variables for ML. Four EFs and six regression-type ML algorithms were tested for *k* estimation accuracy. Each ML model was boosted by adjusting the hyperparameters to get better statistical metrics.

This research is novel because it is the first attempt to use artificial intelligence (AI) for hydraulic conductivity prediction in conjunction with automatic feature selection for Lithuania Quaternary soils. As shown in this study, successful implications of AI techniques should become a common practice while analysing hydrogeological and engineering-geological data.

GEOLOGICAL SETTING

Lithuania is located in northern Europe and is a part of the Baltic Sedimentary Basin, which also covers Latvia, Estonia, parts of Belarus, Poland, Russia and a majority of the Baltic Sea. Surface altitude is from -0.3 to 293 m above sea level (Fig. 1). The dominant highland territory is in south-eastern Lithuania (Medininkai highland) and north-western Žemaičių highland.

Lithuania's climatic conditions are classified as Dfb (D (continental), f (no dry season), b (warm summer)): warm summer, snow, fully humid (Kottek *et al.* 2006). The average temperature is around 7 °C, and the mean annual precipitation is ~ 700 mm (Lithuanian Hydrometeorological Survey 2021).

Quaternary deposits cover the entire Lithuania's ground surface, with common formations and soil types formed during the Pleistocene and Holocene time. Periodic change of climatic conditions during the Quaternary glacial and interglacial epochs formed a variety of soils types, from till to coarse sand and gravel (Šinkūnas, Jurgaitis 1998; Skridlaitė *et al.* 2015). The deposits vary depending on their formation processes and environments; sand and gravel layers have been formed via glacial meltwater flows; till soils were shaped from the material deposited by the glacier; clays and silts have been formed in glacial lakes. Low permeability soils (till soils, clays and silts) cover ~ 60% of Lithuania's surface (Klizas *et al.* 2015).

There are also Pleistocene time aeolian formations (e.g. sand dunes) found in southern Lithuania (Skridlaitė *et al.* 2015). In the river valleys and the Baltic Sea shore, respectively, fluvial and littoral sediment types are found (Skridlaitė *et al.* 2015).

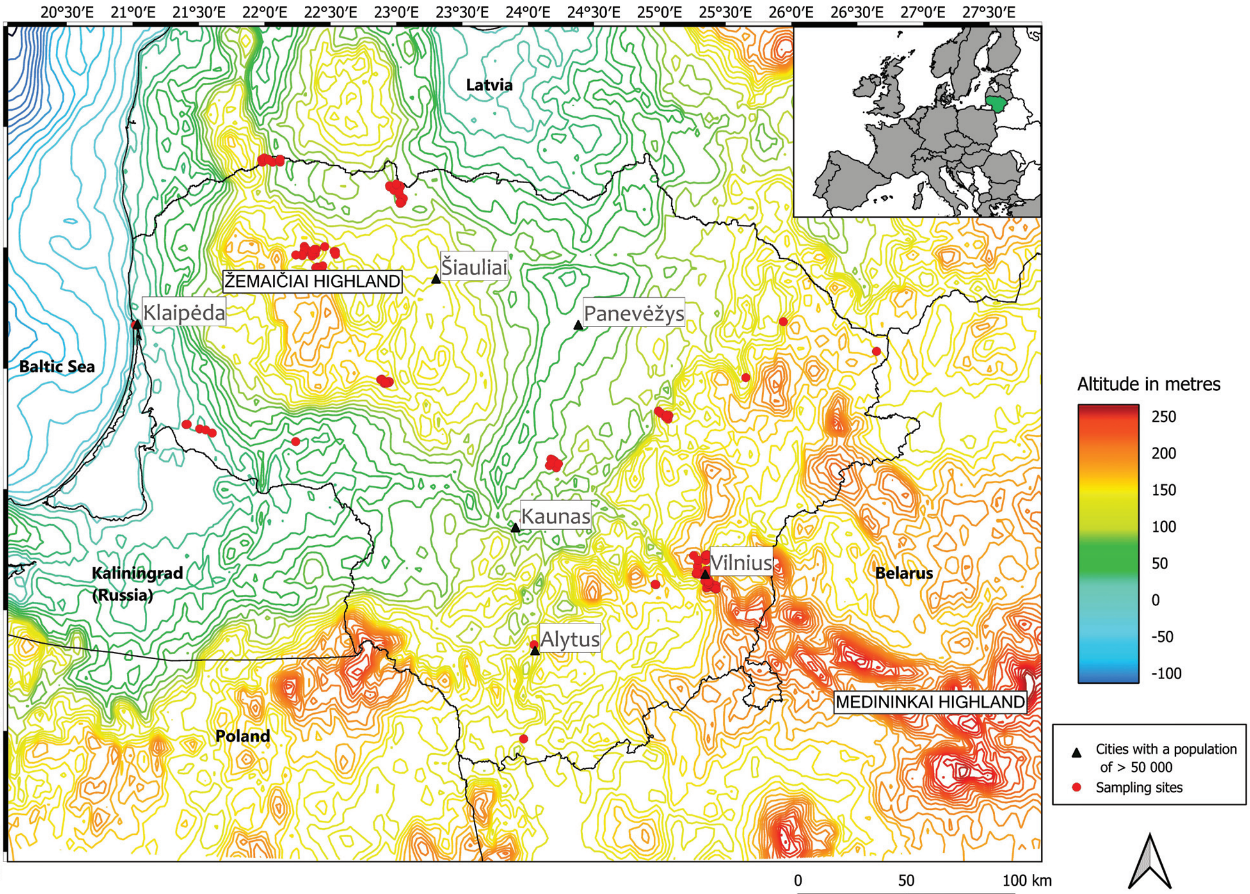

Fig. 1 Sampling site location (red dots), cities (black triangles) and surface elevation model of Lithuania

Quaternary layers‘ thickness varies from 10 to almost 300 m. The average thickness is ~ 130 meters. Northern Lithuania Quaternary deposits are by far the thinnest – from 10 to 30 meters (Guobytė, Satkūnas 2011).

Apart from naturally occurring soils, there is a technogenic soil type that, at some point, was disturbed by human intervention (Skridlaitė *et al.* 2015). It is especially abundant in the urbanized territories where modern and historic anthropogenic influence is vast. A significant amount of data samples used in this research could be attributed to this type, as many sampling sites are clustered in Vilnius and other major cities (Fig. 1). The properties of natural and technogenic soils could be affected by naturally found organic matter (OM). A negative correlation between hydraulic conductivity and OM exists, because the OM present in soil retains water, allowing less water to circulate in porous media (Nemes *et al.* 2005). Contamination with various substances also affects the hydraulic properties of soil (Devatha *et al.* 2019; Xie *et al.* 2024; Zhu *et al.* 2019).

METHODS

Database

The database consists of 282 unique samples developed from laboratory test reports conducted in the Department of Hydrogeology and Engineering Geology at Vilnius University laboratories. The database includes mandatory hydraulic conductivity (k) values, grain size distributions of each sample, coordinates, and sampling depth. Void ratio (e), density (ρ g/cm^3), water content (w), and degree of saturation (S_r) for each sample were also included as complementary data before and after the hydraulic conductivity test. However, the record of these parameters was limited. The samples of this database were collected in fieldwork in the timeframe 2018–2022. The database was stored as a Microsoft Excel spreadsheet file (.xlsx), which is sufficient and versatile for small and medium datasets.

Most samples were collected in cities or their vicinity. A significant part of data came from the western part of Lithuania, where wind turbine parks were developed and geotechnical investigations were conducted (Fig. 1). All samples transported to the laboratory were disturbed.

All hydraulic conductivity testing was done in the laboratory by applying the constant-head method. A total of 239 samples were tested using the International Organization for Standardization (ISO) standard for the constant-head method (ISO 17892-11 2006, 2019), and 43 were tested using a KFZ-type filtrometer (Klizas 2003). The majority of engineering design studies require methods which comply with ISO standards, which is reflected on sample distribution among two methods (Fig. 2). The hydraulic conductivity units in the laboratory reports and the database were expressed in meters per day (m/d). Water used for the test varied in temperature; therefore, all k values were later adjusted to 10 °C, which is a standard reference temperature.

The grain size distribution has been determined using sieving or sedimentation, as described in ISO standard (ISO 146882-4 2016). However, the inconsistency between the obtained fraction intervals obtained in various data entries occurred, and some grain size intervals were unified to get exact matches. The resulting grain size distribution was set to < 0.06, 0.06–0.2, 0.2–0.63, 0.6–2, and > 2 mm. The unification process was validated by adding the sum of each fraction to 100%. Samples that failed validation were discarded.

Grain size diameters (d_{xx} values) are often used for hydraulic conductivity estimation with the EF. In most cases, these values are extracted via visual inspection of the cumulative grain size distribution diagram. Due to a large amount of data in the database, a programming code (using Python script) was developed by the staff of the Department of Hydrogeology and Engineering Geology at Vilnius University to automate the calculation. For this study, d_{xx} values from d_{10} to d_{70} (at an increment of 10%) were acquired by interpolating the cumulative grain size distribution curve (Table 1).

The actual hydraulic conductivity of the data ranges from 0.05 to 27.9 m/d (Table 1). The k intervals (Fig. 2) show that the data distribution is not even, as most samples (87) are in the k range of 0–1 m/d, while the least (31) are > 10 m/d, which may affect the performance of ML models. After the validation of parameters and discarding all entries with limited data, the dataset was shrunk from > 500 to 282 samples. Three samples did not retain the sampling depth (Table 1), but it was still included as the depth was not used as a feature either in the EF or ML. The

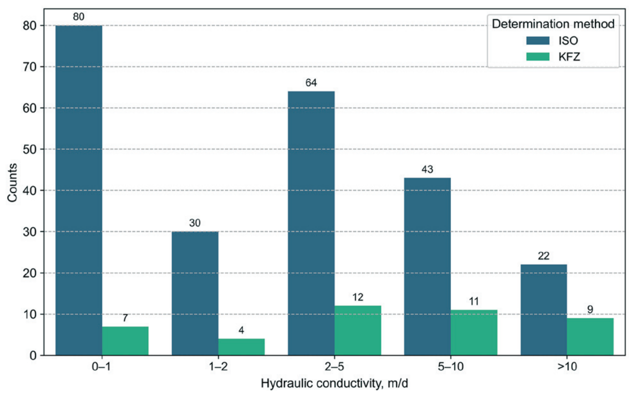

Fig. 2 Sample count of actual hydraulic conductivity interval per determination method from the database

Table. 1 Database parameters' statistical summary for samples tested with ISO and KFZ methods

Statistic		Count	Min	Mean	Max	Q 25%	Q 50%	Q 75%	Std
Depth, m		279	0.20	8.05	27.80	2.85	5.70	11.55	6.69
Grain size, mm	< 0.06	282	0.19	7.01	32.97	2.86	5.06	8.22	6.42
	0.06–0.2	282	1.76	40.88	94.13	17.60	37.97	62.89	25.80
	0.2–0.63	282	0.25	33.25	87.18	18.30	34.25	46.86	19.89
	0.6–2	282	0.01	9.60	57.05	0.77	5.03	15.64	11.10
	> 2	282	0.00	9.26	81.18	0.29	2.67	12.23	14.63
Particle diameter, mm	d_{10}	282	0.03	0.09	0.69	0.06	0.08	0.10	0.07
	d_{20}	282	0.04	0.16	2.04	0.09	0.11	0.18	0.19
	d_{30}	282	0.06	0.23	2.38	0.11	0.15	0.27	0.29
	d_{40}	282	0.08	0.31	2.72	0.13	0.19	0.35	0.38
	d_{50}	282	0.10	0.41	3.06	0.15	0.26	0.44	0.49
	d_{60}	282	0.12	0.54	3.39	0.18	0.35	0.52	0.63
	d_{70}	282	0.14	0.73	4.13	0.20	0.42	0.82	0.82
Hydraulic conductivity k, m/d		282	0.05	4.34	27.90	0.74	2.76	5.91	5.03

sampling depth varied from 0.2 to 27.8 m (average 8.05 m, median (Q 50%) 5.7 m). In this study, grain size and particle diameters d_{10} to d_{70} (Table 1) were selected as features for ML modelling.

Empirical formulas

For over a century, there have been efforts to apply theoretical equations to estimate hydraulic conductivity. Numerous EFs were derived for this purpose using soil grain sizes, porosity, organic matter content, and other parameters (Nemes *et al.* 2005; Odong 2008; Říha *et al.* 2018). In this study, four EFs were tested: Hazen EF (two modifications: coefficient – 100, and custom 40, 100, 140), EF by the United States Bureau of Reclamation (USBR EF), and Alyamani and Sen EF (Alyamani, Şen 1993; Hazen 1892; Klizas 2003; Odong 2008; Urumović *et al.* 2020).

Allen Hazen developed an EF for determining the hydraulic conductivity of saturated sands using effective particle size d_{10} (Equation 1) (Hazen 1892; Klizas 2003). The basic form of the equation uses the default empirical coefficient *C* as 100. However, *C* varies depending on the soil type and may range from 1–1000 (Carrier 2003). For example, *C* 46 is a better fit for clayey sand, while 142 for coarse sand (Klizas 2003). In this study, hydraulic conductivity was estimated using the Hazen formula using empirical coefficient 100 (abr. HS – Hazen Simple) and attributing three distinct values (40, 100, 140) based on the soil type (abr. HC – Hazen Custom): 40 for silty and clayey soil,140 for sand with a high content of gravel, and 100 for the rest of soil types. Only the samples that met conditions $d_{60}/d_{10} \leq 5$ and $d_{10} < 3.0$ mm were used. In Equation 1, *k* is hydraulic conductivity in cm/s, *C* is the dimensionless empirical coefficient, and d_{10} is the effective grain size in centimetres.

$$k = C \times d_{10}^{2} \tag{1}$$

The USBR EF utilizes grain size d_{20} in the determination of k (Equation 2) (Urumović *et al.* 2020). The database was filtered to select entries where $d_{60}/d_{10} \leq 5$. In Equation 2, *k* is expressed in cm/s, and d_{20} in centimetres. For further analysis, both Hazen and USBR, *k* units were converted to m/d.

$$k = 0.36 \times d_{20}^{2.3} \tag{2}$$

The EF developed by Alyamani and Şen (AS) (Equation 3) uses two particle diameter values for *k* estimation (Alyamani, Şen 1993; Odong 2008). I_0 represents the intercept (mm) of the line created by plotting d_{50} and d_{10} against the grain-size axis. Here, d_{10} denotes the effective grain diameter (mm), while d_{50} signifies the median grain diameter (mm) (Odong 2008). Hydraulic conductivity k is expressed in m/d.

$$k = 1300 \times (I_0 + 0.025\,(d_{50} - d_{10}))^2 \tag{3}$$

These equations were used to estimate hydraulic conductivity values for the database entries. The determination coefficient (R^2) and the Mean Absolute Error (MAE) were calculated for each EF's result.

Machine learning

The ML modelling was executed on Jupyter Notebook using Python version 3.9.12. Data processing, cleaning, and EF calculation were made utilizing numpy (Harris *et al.* 2020) and pandas (McKinney 2010) libraries. Bar charts, Taylor diagram, and correlation plots were compiled using the matplotlib library (Hunter 2007). ML algorithms provided in the scikit-learn library were used for modelling (Pedregosa *et al.* 2011).

In this study, ML regression algorithms were used as predictive models to analyze and estimate continuous data outcomes based on input data. This approach is part of supervised learning, where an algorithm is trained on a labelled dataset containing both input features (independent variables) and their corresponding target values (dependent variables). During modelling, patterns and relationships are identified in the data. Trained models can then be applied to new, unseen data. Considering hydraulic conductivity is a continuous value – regression type ML algorithms were used in this study.

Six different algorithms were used in this study attempting to estimate *k* values:

1. ElasticNet (EN) is a linear regression model that includes options for L1 and L2 regularization. This model is particularly useful when dealing with multiple correlated features, as it facilitates both feature selection and coefficient shrinkage (Hao *et al.* 2020; Yasin *et al.* 2022).
2. The Gradient Boosting Regressor (GBR) constructs an ensemble of decision trees in a sequential manner, where each tree aims to correct the errors made by the preceding ones. By focusing on minimizing the loss function, this model enhances performance and proves to be highly effective for a variety of regression tasks (Natekin, Knoll 2013; Zeleke *et al.* 2024).
3. The Huber Regressor (HR) is a variant of linear regression designed to minimize the effect of outliers. It uses different loss functions for different error magnitudes: squared loss for small errors and absolute loss for larger errors. This dual approach diminishes the impact of outliers while maintaining accuracy for normally distributed data (Hao *et al*. 2020; Yasin *et al*. 2022).
4. The K-neighbors Regressor (KN) predicts the target value by averaging the values of the closest data points (*K*) in the feature space. This non-parametric method is effective for capturing local data patterns and relationships, making it versatile for various regression problems (Araya, Ghezzehei 2019; Motevalli *et al.* 2019; Zeleke *et al.* 2024).
5. Multi-layer Perceptron Regressor (MLPR) is a neural network-based regression model that constructs multiple layers of interconnected neurons to learn complex patterns in data. By leveraging multiple hidden layers, this method captures intricate relationships and interactions within the data, enhancing predictive accuracy and flexibility for a variety of regression tasks (Arshad *et al.* 2013; Huang *et al.* 2019).
6. The Random Forest Regressor is an ensemble technique that builds several decision trees in parallel during training and aggregates their predictions to boost accuracy. By combining multiple trees, this method effectively addresses overfitting and enhances both predictive performance and robustness (Natekin, Knoll 2013; Zeleke *et al.* 2024).

Prior to modelling, input data was separated into twelve features (grain size and particle diameter) and the target (hydraulic conductivity) (Fig. 3, Input data structure).

An automated process was programmed to create a list of all possible combinations of twelve input features, e.g. *('GS < 0.06', 'GS_0.2–0.63', 'GS_0.6–2',* *'d_{10}', 'd_{40}', 'd_{70}'), ('GS_0.6–2', 'd_{10}', 'd_{40}', 'd_{60}'), ('d_{10}', 'd_{20}', 'd_{30}')*, resulting in 4095 unique variants. All feature values were standardized using the Standard Scaler method from the scikit-learn library. Standardization results in every feature having a mean of 0, and the standard deviation is equal to one. This ensures a proper training of the ML algorithm as all features have the same structure.

The dataset was divided into training and testing sets, with 25% of the data reserved for testing. The samples were randomly assigned to the train set (211 samples) and the test set (71 samples).

A cross-validation technique was applied to the train set to ensure that ML modelling results were reliable and not coincidental. The train set was split into five random subsets. R^2 and MAE were calculated for each partition. The final R^2 and MAE of the modelling are an average of all five splits.

The initial ML step encompassed the modelling of every 4095 feature combination to each algorithm (EN, HR, RFR, KNR, GBR, MLPR), with the aim to test all possible feature combinations without any initial presumptions of which might be more significant. The data for each feature combination, model, resulting R^2, and MAE for test and train sets were collected in the intermediate results table (Fig. 3, Initial ML).

The results were processed by omitting the entries where the standard deviation of a train set R^2_{SD} was below 0.08 (Fig. 3, Feature combination selection) to keep models that perform consistently during cross-validation. The value 0.08 was estimated based on the ML results, where $R^2 \sim 0.4$ should not exceed 20%. Ideally, these scores should match, but in practice, the model typically performs slightly worse on unseen data (the test set) compared to the training data. A significant discrepancy between the two indicates overfitting, meaning the model has learned the training data too intricately and performs poorly on new data.

The models with a difference in R^2 scores between the train and test sets below 0 or above 0.1 were discarded to ensure no marginal overfitting was left (Fig. 3, Feature combination selection). The remaining models were sorted to estimate the best performers, resulting in 29 feature combinations that matched the conditions discussed. In this step, only combinations were selected; the algorithms that yielded them were ignored.

The selected 29 combinations were passed for modelling again. During this step, ML algorithms

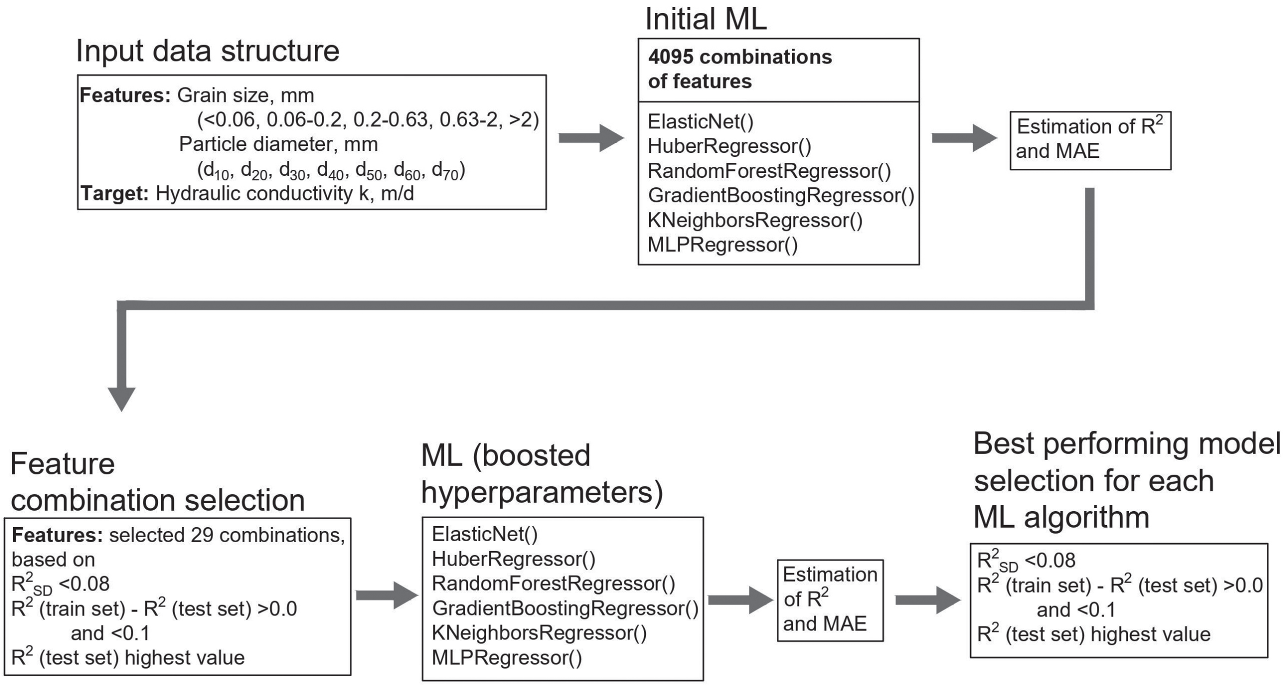

Fig. 3 ML study design scheme

were tuned using a grid-search approach (Fig. 3, ML (boosted hyperparameters)). Each algorithm has a range of hyperparameters that control the ML training process. Boosting the hyperparameters may result in a better-performing model. The hyperparameters of each algorithm were selected to adjust the resulting models' robustness, precision, and sensitivity and to deal with overfitting. The grid-search technique systematically searches through a predefined set of hyperparameters to identify the tuning that yields the best performance. The technical description of each model's hyperparameters and their application are provided in Supplemental Material.

Tuning results were collected into a new dataframe. The data was again processed, and entries where $R^2_{SD} < 0.08$ and R^2 scores difference between the train and test sets > 0 and < 0.1 were selected. The remaining results were sorted by the highest value of the R^2 test set. The best-performing combination of features and the optimal hyperparameter set for each of the ML algorithms were selected as the final result set. This set includes MAE and R^2 of the test set, train set and standard deviation obtained during the cross-validation of the train set.

RESULTS AND DISCUSSION

The data used in this study proved to be difficult to model due to a variety of soil types which was applied to modelling. The results of both ML and EF are moderate, with the highest R^2 being 0.46 and the lowest 0.10 (Figs 4, 5; Table 2). Hydraulic conductivity estimations' mean absolute error ranged between 2.31–6.54 m/d (Figs 4, 5; Table 3). The R^2 and MAE values of ML discussed in this chapter are for the test set, which shows realistic metrics on how the model performed.

The R^2 of ML models varies from 0.39 (KNR) to 0.46 (MLPR), while the EF ranges from 0.1 (AS) to 0.33 (HS). Accordingly, the MAE for ML is 2.31 (EN) to 2.81 m/d (RFR), and the EF varies from 3.05 (USBR) to 6.54 m/d (AS) (Tables 2, 3).

ML models outperformed all EFs based on R^2 and MAE. Taylor's diagram, which plots the statistical metrics (radius MAE, angle R^2) of ML and EF groups, clearly shows the advantages of ML models (Fig. 5). The determination coefficient of ML models is higher than of EFs, and MAE is up to twice more accurate.

The results are similar to those obtained during the modelling of ML and EF with the manual selection of features, with a smaller dataset (Vanhala 2024). The best-performing model R^2 ML – 0.47, EF – 0.38, and the lowest obtained MAE for the ML model was 2.28 (Vanhala 2024).

Hydraulic conductivity estimation research on Lithuanian soil yielded high determination coefficients (R^2 up to 0.96) (Dobkevičius 2002), which is significantly higher than the ML and EF attempts of this study. However, the study described in Dobkevičius (2002) deals with a significantly smaller dataset (48 samples) in which samples are mainly selected from a few sites, which could lead to a bias towards certain soil types or other circumstances, such as organic matter content. Samples for this study are collected from over 100 sites across Lithuania. A sig-

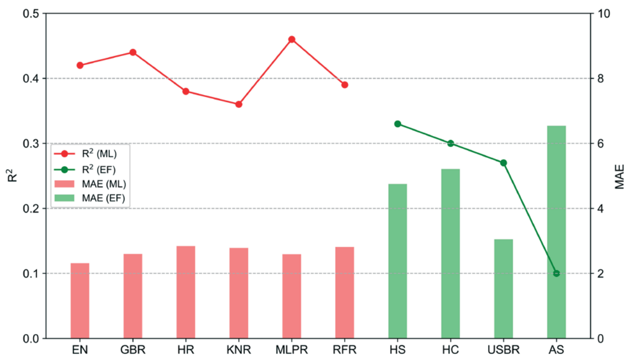

Fig. 4 R^2 and MAE bar chart of ML modelling and EF calculation results

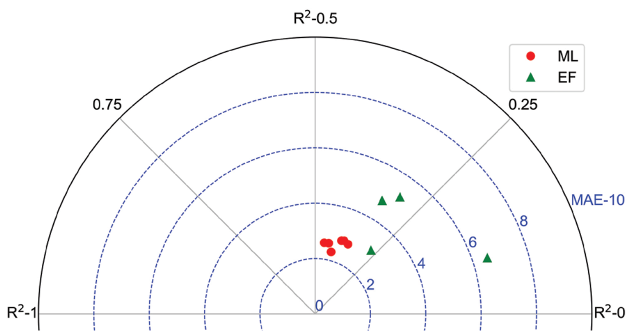

Fig. 5 Taylor diagram of R^2 (angular coordinate) and MAE (radial distance) of ML modelling and EF calculation results

nificant part of hydraulic conductivity determination took place *in situ* (contrary to exclusive lab testing in this study), which is also a major factor considering theoretical and actual *k* correlation.

The results and comparison with other similar studies pose a question about the optimal dataset size required for a proper estimation of *k* in the Quaternary deposits of Lithuania. The modelling of small datasets consisting of a few dozen samples may lead to an almost ideal correlation. The data used in this study include a couple hundred samples, which are diverse and collected on different sites but yield only moderate correlation. A study with over 18 thousand samples (4661 for the test set) managed to get an overall R^2 equal to 0.90 (Araya, Ghezzehei 2019), which is still less than obtained by (Dobkevičius 2002). However, each soil type modelled separately yielded different correlations, e.g. sandy clay (n – 89, R^2 – 0.497), loamy clay (n – 443, R^2 – 0.844), silt (n – 6, R^2 – 0.968), sand (n – 2951, R^2 – 0.875) (Araya, Ghezzehei 2019). This clearly outlines that only a few entries, as well as a couple of thousands of the same soil type, could result in high statistical metrics after ML modelling.

The current Lithuania soil dataset is too small to

Table 2 R^2 of ML performance for train and test sets for the best fitting feature combination of an algorithm with boosted hyperparameters. SD – standard deviation. EF results are represented as R^2

Model	Best fitting feature combination	R^2 (test set)	R^2 (train set)	R^2 (SD)
EN	['GS < 0.06', 'GS_0.06–0.2', 'd$_{10}$', 'd$_{30}$', 'd$_{70}$']	0.42	0.43	0.06
GBR	['GS < 0.06', 'GS_0.2–0.63', 'GS_0.6–2', 'd$_{10}$', 'd$_{40}$', 'd$_{70}$']	0.44	0.46	0.07
HR	['GS < 0.06', 'GS_0.06–0.2', 'GS_0.2–0.63', 'GS_0.6–2', 'GS_> 2', 'd$_{10}$', 'd$_{30}$', 'd$_{40}$']	0.38	0.46	0.07
KNR	['GS_0.06–0.2', 'd$_{10}$', 'd$_{30}$', 'd$_{70}$']	0.36	0.45	0.04
MLPR	['GS < 0.06', 'd$_{10}$', 'd$_{60}$', 'd$_{70}$']	0.46	0.50	0.09
RFR	['GS < 0.06', 'GS_> 2', 'd$_{10}$', 'd$_{30}$', 'd$_{40}$', 'd$_{70}$']	0.39	0.42	0.06
Empirical formula	–	R^2	–	–
HS	–	0.33	–	–
HC	–	0.30	–	–
USBR	–	0.27	–	–
AS	–	0.10	–	–

Table 3 MAE of ML performance for train and test sets for the best fitting feature combination of an algorithm with boosted hyperparameters. SD – standard deviation. EF results are represented as MAE

Model	Best fitting feature combination	MAE (test set)	MAE (train set)	MAE (SD)
EN	['GS < 0.06', 'GS_0.06–0.2', 'd$_{10}$', 'd$_{30}$', 'd$_{70}$']	2.31	2.75	0.14
GBR	['GS < 0.06', 'GS_0.2–0.63', 'GS_0.6–2', 'd$_{10}$', 'd$_{40}$', 'd$_{70}$']	2.60	2.50	0.17
HR	['GS < 0.06', 'GS_0.06–0.2', 'GS_0.2–0.63', 'GS_0.6–2', 'GS_> 2', 'd$_{10}$', 'd$_{30}$', 'd$_{40}$']	2.84	2.35	0.13
KNR	['GS_0.06–0.2', 'd$_{10}$', 'd$_{30}$', 'd$_{70}$']	2.78	2.58	0.17
MLPR	['GS < 0.06', 'd$_{10}$', 'd$_{60}$', 'd$_{70}$']	2.59	2.35	0.23
RFR	['GS < 0.06', 'GS_> 2', 'd$_{10}$', 'd$_{30}$', 'd$_{40}$', 'd$_{70}$']	2.81	2.60	0.16
Empirical formula	–	MAE	–	–
HS	–	4.75	–	–
HC	–	5.21	–	–
USBR	–	3.05	–	–
AS	–	6.54	–	–

fully realize the potential of ML approach. The dataset includes 32 unique soil types, and a robust ML model for Lithuania's Quaternary soils would require at least a few hundred distinct entries for each type. While some soil types could be consolidated into groups, the complete dataset still needs to comprise thousands of samples to be effective.

However, the database is sufficient for comparing the results obtained with the EF and ML, both of which were tested against the same conditions. The previously discussed differences in statistical metrics could also be supported by visually analysing correlation diagrams (Figs 6, 7).

EF (Fig. 6) and ML (Fig. 7) correlation plots represent the distribution of actual and predicted hydraulic conductivity values, complemented with a trendline. The correlation line shows an ideal match between predicted and actual values. In ML calibration plots, the trendline is derived from test set data.

Both Hazen EF modifications show that the *k* estimations made are higher than the actual (significantly more data points and the trendline are above the calibration line) (Fig. 6, a, b). Visual inspection reveals that a minority of data points cluster around the calibration line; the correlation is weak.

The USBR EF diagram has more data points distributed under the correlation line, meaning that the model results in predictions that are less than actual (Fig. 6, c). The trendline is close to the calibration line and almost parallel, but a small R^2 shows poor correlation as well.

The Alyamani and Sen EF performed the poorest of all models tested, showing almost no correlation, despite using two soil parameters (d_{10} and d_{50}), while Hazen and USBR EFs accepted only one – d_{10} and d_{20}, respectively.

The Hazen EF performed the best and could be improved by adjusting empirical coefficient C spe-

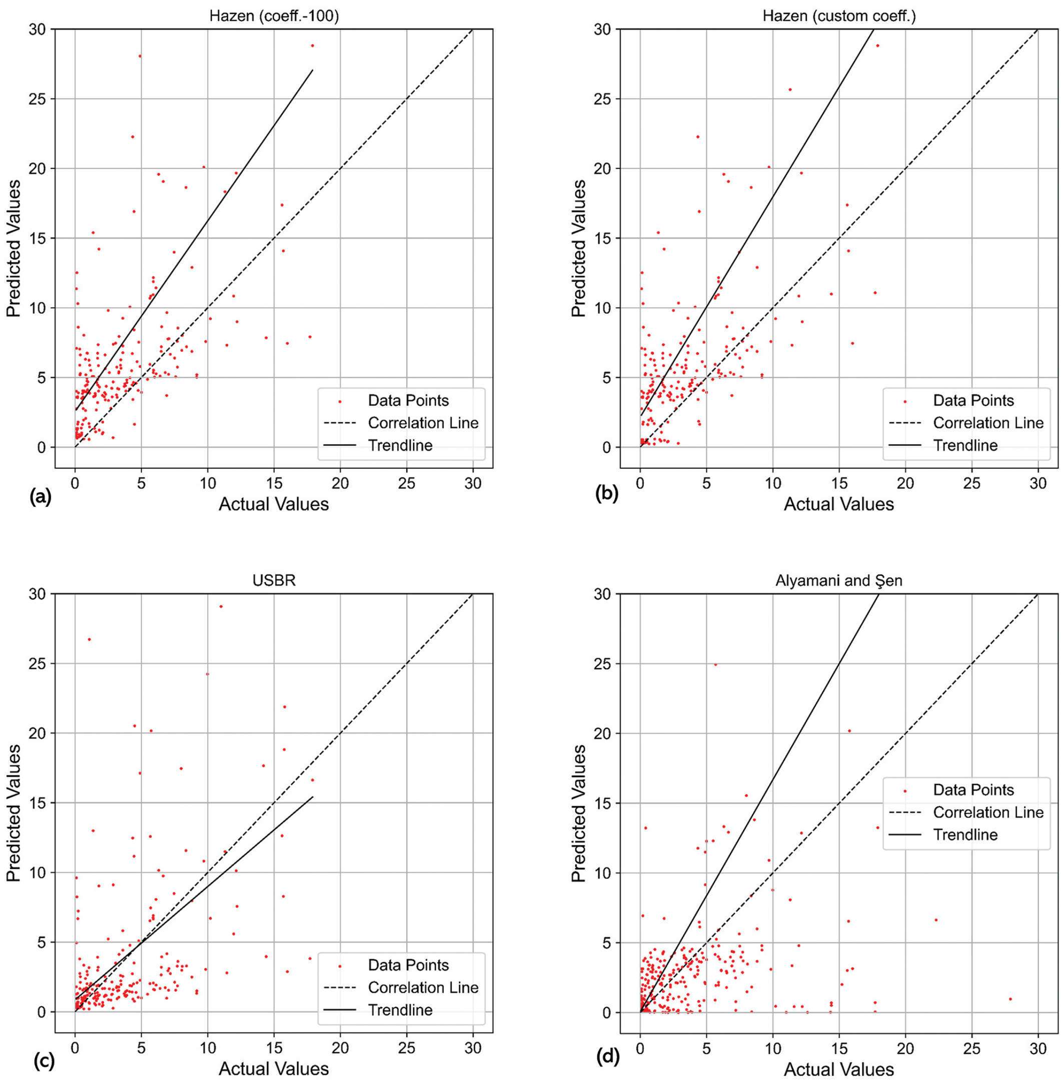

Fig. 6 EF correlation diagrams. Correlation plots of hydraulic conductivity predicted *vs*. actual values: a) Hazen Simple, b) Hazen Custom, c) USBR, d) Alyamani and Şen. The *dashed line* shows ideal correlation, and the *black solid line* is the trendline

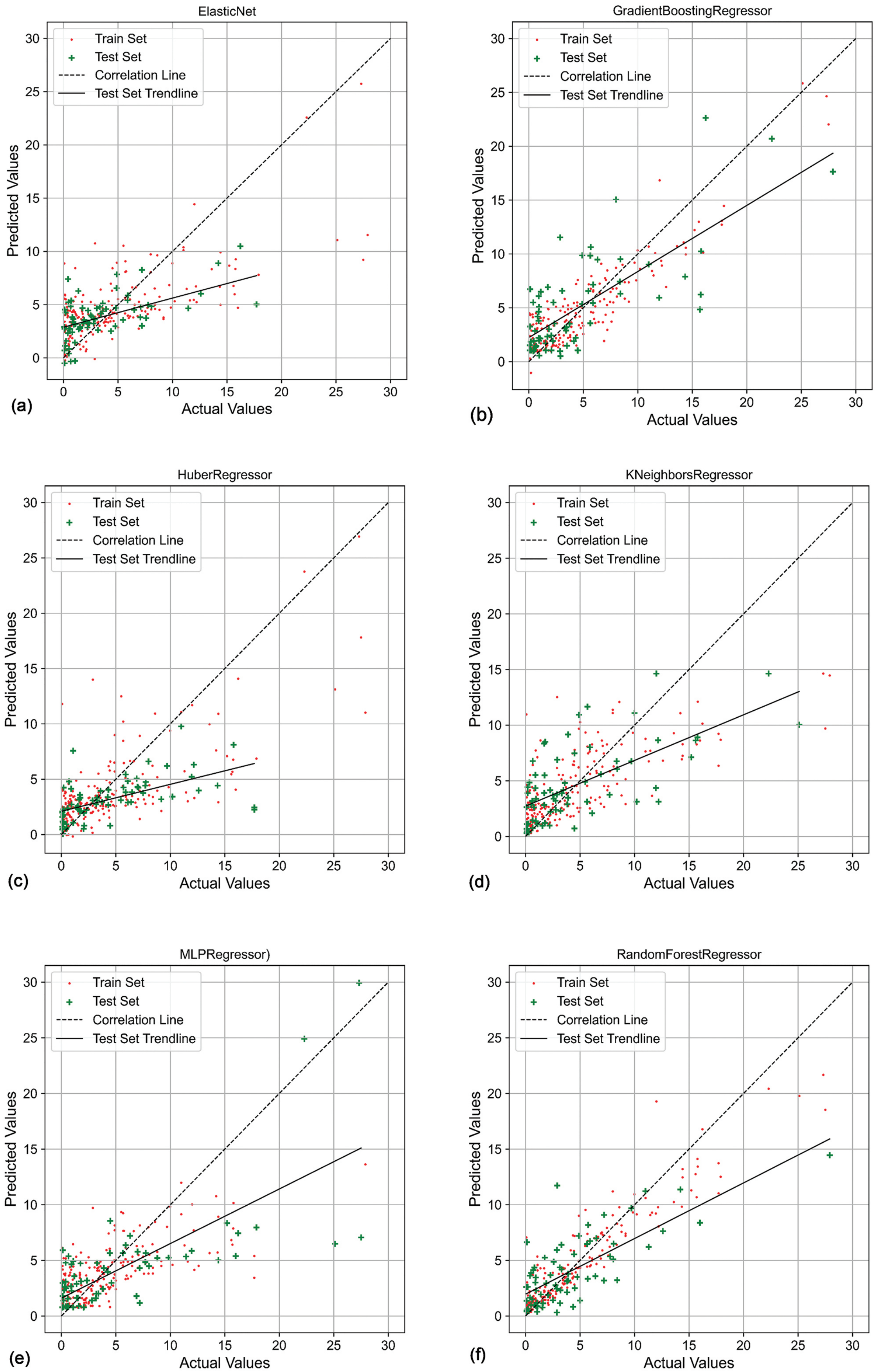

Fig. 7 ML correlation diagrams Correlation plots of hydraulic conductivity predicted *vs*. actual values: a) Elastic Net, b) Gradient Boosting, c) Huber, d) K-neighbours, e) MLP, f) RandomForest. The dashed line shows ideal correlation, and the black solid line is the trendline

cifically for Lithuanian soil types or by adding additional limitations on soil grain size distribution or d_{xx}. Also, a specific EF could be derived for each soil type or k value, as was attempted by previous research (Dobkevičius 2002). However, ML algorithms are more accurate, and future research should be focused on developing this field.

The visual inspection of ML calibration graphs shows that GBR and RFR models have a clear data point (both train and test sets) clustering around the calibration line (Fig. 7, b, f). However, the GBR test set data points are way more scattered than the test set, indicating overfitting (Fig. 7, b). The data of the RFR model test and train sets are distributed more evenly, meaning that the model performs well on the unseen data (Fig. 7, f). The GBR model worked best with feature combination (*'GS < 0.06', 'GS_0.2–0.63', 'GS_0.6–2',* 'd_{10}', 'd_{40}', 'd_{70}'), test set R^2 – 0.44 (Table 2). The RFR model performed utilizing combination (*'GS < 0.06', 'GS_> 2',* 'd_{10}', 'd_{30}', 'd_{40}', 'd_{70}'), test set R^2 – 0.39 (Table 2).

The EN estimator, based on linear type regression with normalization options, performed best with feature combination (*'GS < 0.06', 'GS_0.06–0.2',* 'd_{10}', 'd_{30}', 'd_{70}'), yielding the test set R^2 of 0.42 (Table 2). The calibration plot shows a similar distribution between test and train sets, suggesting little overfitting, but data points do not tend to cluster around the calibration line evenly (Fig. 7, a).

The HR model, based on linear regression designed to handle outliers, performed best with feature combination (*'GS < 0.06', 'GS_0.06–0.2', 'GS_0.2–0.63', 'GS_0.6–2', 'GS_> 2',* 'd_{10}', 'd_{30}', 'd_{40}'), achieving the test set R^2 of 0.38 (Table 2). The calibration plot indicates a similar distribution between test and train sets, suggesting minimal overfitting. However, the data points do not cluster evenly around the calibration line (Fig. 7, c). This model also incorporated the most features (8) compared to other ML models (4–6).

The KNR algorithm, which operates by averaging the values of the nearest data points, yielded the poorest performance among all evaluated ML models. The optimal feature combination (*'GS_0.06–0.2',* 'd_{10}', 'd_{30}', 'd_{70}') produced a test set R^2 of 0.38 and a train set R^2 of 0.46 (Table 2). Despite the trendline showing a noticeable deviation from the calibration line, the extent of overfitting remains minimal (Fig. 7, d), indicating the model's limited predictive capability but acceptable generalization behaviour.

The MLPR algorithm, based on a neural network, performed the best among all ML models using feature combination (*'GS < 0.06',* 'd_{10}', 'd_{60}', 'd_{70}'), yielding the test set R^2 of 0.46 and train set R^2 of 0.50 (Table 2). Although the trendline deviates from the calibration line, minimal overfitting is observed (Fig. 7, e).

In all cases, the test set data trendline is below the calibration line for k values > 5 m/d, and above when $\underline{k}$ < 5 m/d, meaning that ML models are biased by predicting higher values for lower permeability samples and lowering estimations for a higher k.

The feature combinations selected for ML included 4–8 different parameters, yielding better results compared to the EF, which only uses one or two parameters. All combinations chosen for the final ML step share a common d_{10} feature, which is also utilized in the Hazen and Alyamani and Şen formulas. Grain size (GS) and d_{xx} values are included in all combinations at least once, indicating their necessity for optimal modelling in hydraulic conductivity estimation. The results from feature selection highlight the importance of optimizing input parameters when estimating k – more features do not necessarily improve model performance. None of the models performed efficiently with all available features, and the minimum number of features required was four.

This study focused on modelling hydraulic conductivity based on grain size alone. Other common soil parameters such as porosity, organic matter, mineralogical composition, and density were omitted from the feature list as there was little data available on these instances. Having these parameters in a database in substantial amounts would likely help future research on AI applications for hydraulic conductivity determination. The other important notice is that ML modelling should be conducted with a sufficient database size, accounting for the diversity of the Lithuanian Quaternary deposit variety. A clear advantage of ML against the EF shown in this study should be considered when dealing with other similar scientific and applied problems in hydrogeology and engineering geology. A majority of previous Lithuanian studies in these fields should be revised and updated using modern computational techniques.

CONCLUSIONS

The ML models outperformed EFs in estimating hydraulic conductivity. Among the four EFs evaluated (Hazen with a constant coefficient of 100; Hazen with custom coefficients of 40, 100, and 140; USBR; and Alyamani and Şen), the Hazen equation with a custom coefficient of 100 performed the best, achieving an R^2 of 0.33. In comparison, the ML models provided more accurate predictions with R^2 values ranging from 0.36 to 0.46.

Despite the ML models' superior performance, the overall modelling results were moderate (R^2 < 0.50), primarily due to the insufficient size of the database and a wide range of soil types. The development of a robust general model for all Lithuanian Quaternary soil types would require a database comprising several hundred unique samples per soil type. In contrast, this study's dataset included 282 samples spanning 32 soil types. Additionally, the incorporation of parameters such as porosity, organic matter content, and soil density could enhance the training of ML algorithms.

Each ML algorithm achieved optimal performance with a customized combination of input parameters, specifically grain size distribution and particle sizes, ranging from 4 to 8 features. The use of either too few or too many features resulted in diminished performance.

Future studies on this topic should account for the increasing database size, adding more features, such as particles‘ mineralogical content, organic matter, contamination traces (e.g. contaminated or not), and genesis (glacial, fluvial, etc.).

ML algorithms significantly improved the accuracy of hydraulic conductivity estimations compared to EFs. The use of effective ML models for k estimation may broaden any geological database where other soil parameters are available and be a tool for synthetic data augmentation. Thus, less reliable than actually measured, the modelled hydraulic conductivity could be used for mapping or other spatial investigations of various areas and regions and could be applied for urban planning, environmental assessments and runoff estimations. Apart from that, the results of this study suggest that AI has substantial potential for broader application in hydrogeological and engineering geology research and practical implementations.

ACKNOWLEDGEMENTS

The authors would like to express sincere gratitude to the two anonymous reviewers for their insightful feedback and constructive suggestions, which have significantly enhanced the quality of this manuscript. The authors also extend their deepest appreciation to journal editors for their professional handling of the review process and thoughtful guidance.

REFERENCES

Alyamani, M.S., Şen, Z. 1993. Determination of Hydraulic Conductivity from Complete Grain-Size Distribution Curves. *Groundwater 31*(*4*), 551–555. https://doi.org/https://doi.org/10.1111/j.1745-6584.1993.tb00587.x

Araya, S.N., Ghezzehei, T.A. 2019. Using Machine Learning for Prediction of Saturated Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. *Water Resources Research 55*(*7*), 5715–5737. https://doi.org/10.1029/2018WR024357

Arshad, R.R., Sayyad, Gh., Mosaddeghi, M., Gharabaghi, B. 2013. Predicting Saturated Hydraulic Conductivity by Artificial Intelligence and Regression Models. *ISRN Soil Science* 1–8. https://doi.org/10.1155/2013/308159

Bhaurao More, S., Deka, P.C., Prakash Patil, A., Naganna, S.R. 2022. Machine learning-based modeling of saturated hydraulic conductivity in soils of tropical semi-arid zone of India. *Sådhanå*, *47*(*26*). https://doi.org/10.1007/s12046-022-01805-6S

Carrier, W.D. 2003. Goodbye, Hazen; Hello, Kozeny-Carman. *Journal of Geotechnical and Geoenvironmental Engineering 129*, 1054–1056.

Devatha, C.P., Vishnu Vishal, A., Purna Chandra Rao, J. 2019. Investigation of physical and chemical characteristics on soil due to crude oil contamination and its remediation. *Applied Water Science 9*(4). https://doi.org/10.1007/s13201-019-0970-4

Dobkevičius, M. 2002. *Hydraulic Conductivity Determination Methods' Research and Improvement in Lithuania Vadose Zone Soils*. Vilnius University, 75 pp. [In Russian].

Guobytė, R., Satkūnas, J. 2011. Chapter 19 – Pleistocene Glaciations in Lithuania. In: Ehlers, J., Gibbard, P.L., Hughes, P.D. (Eds), *Quaternary Glaciations – Extent and Chronology* 15, 231–246. Elsevier. https://doi.org/10.1016/B978-0-444-53447-7.00019-2

Hao, X., Zhao, Y., Wang, Y. 2020. Forecasting the real prices of crude oil using robust regression models with regularization constraints. *Energy Economics 86*. https://doi.org/10.1016/j.eneco.2020.104683

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E. 2020. Array programming with NumPy. *Nature 585*(*7825*), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hazen, A. 1892. Some Physical Properties of Sands and Gravels, with Special Reference to Their Use in Filtration. *24th Annual Report, Massachusetts State Board of Health*, 539 pp. Massachusetts State Board of Health.

Huang, X., Gao, L., Crosbie, R.S., Zhang, N., Fu, G., Doble, R. 2019. Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning. *Water* (*Switzerland*) *11*(9). https://doi.org/10.3390/w11091879

Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ibrahim, M., Aliyu, J. 2016. Comparison of Methods for Saturated Hydraulic Conductivity Determination: Field, Laboratory and Empirical Measurements. *British Journal of Applied Science & Technology 15*(3), 1–8. https://doi.org/10.9734/bjast/2016/24413

ISO 17892-11. 2006. *Geotechnical investigation and testing – Laboratory testing of soil – Part 11. Permeability tests (ISO 17892-11): 2006.*

ISO 146882-4. 2016. *Geotechnical investigation and testing – Identification and classification of soil. Part 4:*

Determination of particle size distribution (ISO 146882-4: 2016).

ISO 17892-11. 2019. *Geotechnical investigation and testing – Laboratory testing of soil – Part 11. Permeability tests (ISO 17892-11): 2019*.

Jang, J., Narsilio, G.A., Santamarina, J.C. 2011. Hydraulic conductivity in spatially varying media-a pore-scale investigation. *Geophysical Journal International 184*(*3*), 1167–1179. https://doi.org/10.1111/j.1365-246X.2010.04893.x

Juodkazis, V., Gregorauskas, M., Mokrik, R. 2012. Regioninė Hidrogeodinamika [Regional Hydrogeodynamics]. Vilnius University Press, 248 pp. [In Lithuanian].

Klimašauskas, M., Šaulys, V., Baublys, R., Survilė, O. 2020. Hydraulic Conductivity of Drainage Ditch Backfill with a Lime Additive in Clay Soils. *Environmental Engineering and Management Journal 18(3)*, 497–504.

Klizas, P. 2003. Hidrogeologijos laboratoriniai darbai [Hydrogeological laboratory works]. Vilnius University Publishing House, 111 pp. [In Lithuanian].

Klizas, P. 2014. Geofiltration Studies of Clay at the Future Radioactive Waste Repository for Ignalina Nuclear Power Plant. *Journal of Environmental Engineering and Landscape Management 20*, 1–7. https://doi.org/10.3846/16486897.2014.903186

Klizas, P., Šečkus, R. 2007. Filtration and geoelectrical investigations in the karst region of North Lithuania. *Geologija 59*, 77–81.

Klizas, P., Gadeikis, S., Žilionienė, D. 2015. Evaluation of moraine loams' filtration properties. *The Baltic Journal of Road and Bridge Engineering 10*(*4*), 293–298. https://doi.org/10.3846/bjrbe.2015.37

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F. 2006. World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift 15*(*3*), 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Li, Y., Rahardjo, H., Satyanaga, A., Rangarajan, S., Lee, D.T.T. 2022. Soil database development with the application of machine learning methods in soil properties prediction. *Engineering Geology 306*. https://doi.org/10.1016/j.enggeo.2022.106769

Lithuanian Hydrometeorological Survey. 2021. Lietuvos klimato pokyčių vertinimas lyginant 1961–1990 m. ir 1991–2020 m. standartines klimato normas [Comparative evaluation of climate changes in Lithuania 1961–1990 and 1991–2020 standard climate norms], 18 pp. [In Lithuanian].

McKinney, W. 2010. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference 1*(*Scipy*), 56–61. https://doi.org/10.25080/majora-92bf1922-00a

Motevalli, A., Naghibi, S.A., Hashemi, H., Berndtsson, R., Pradhan, B., Gholami, V. 2019. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. *Journal of Cleaner Production 228*, 1248–1263. https://doi.org/10.1016/j.jclepro.2019.04.293

Natekin, A., Knoll, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics 7*(*DEC*). https://doi.org/10.3389/fnbot.2013.00021

Nemes, A., Rawls, W.J., Pachepsky, Y.A. 2005. Influence of Organic Matter on the Estimation of Saturated Hydraulic Conductivity. *Soil Science Society of America Journal 69*(*4*), 1330–1337. https://doi.org/10.2136/sssaj2004.0055

Odong, J. 2008. Evaluation of Empirical Formulae for Determination of Hydraulic Conductivity based on Grain-Size Analysis. *The Journal of American Science*, *4*(*1*).

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, É. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*(*85*), 2825–2830.

Rehman, Z. ur, Khalid, U., Ijaz, N., Mujtaba, H., Haider, A., Farooq, K., Ijaz, Z. 2022. Machine learning-based intelligent modeling of hydraulic conductivity of sandy soils considering a wide range of grain sizes. *Engineering Geology 311*. https://doi.org/10.1016/j.enggeo.2022.106899

Říha, J., Petrula, L., Hala, M., Alhasan, Z. 2018. Assessment of empirical formulae for determining the hydraulic conductivity of glass beads. *Journal of Hydrology and Hydromechanics 66*(*3*), 337–347. https://doi.org/10.2478/johh-2018-0021

Skridlaitė, G., Prušinskienė, S., Jakimavičiūtė-Maselienė, V., Gadeikytė, S. 2015. *Engineering Geology*. VGTU, Technika. [In Lithuanian]. https://doi.org/10.3846/1509-S

Šinkūnas, P., Jurgaitis, A. 1998. *Ledyninių nuogulų litologija ir sedimentacija.* Vilnius: Academia, 71 pp.

Urumović, K., Borović, S., Urumović, K., Navratil, D. 2020. Validity range and reliability of the United States Bureau of Reclamation (USBR) method in hydrogeological investigations. *Hydrogeology Journal 28*(*2*), 625–636. https://doi.org/10.1007/s10040-019-02080-2

Vanhala, E.K.M. 2024. *Theoretical Hydraulic Conductivity Determination of Lithuanian Soil Samples* [Master Thesis]. Vilnius University.

Veloso, M.F., Rodrigues, L.N., Filho, E.I.F. 2022. Evaluation of machine learning algorithms in the prediction of hydraulic conductivity and soil moisture at the Brazilian Savannah. In: *Geoderma Regional 30*. https://doi.org/10.1016/j.geodrs.2022.e00569

Xie, Y., Wang, H., Guo, Y., Wang, C., Cui, H., Xue, J. 2024. Effects of microplastic contamination on the hydraulic, water retention, and desiccation crack properties of a natural clay exposed to leachate. *Journal of Environmental Management 351*. https://doi.org/10.1016/j.jenvman.2023.119858

Yasin, B., Maraqa, F., Al-Sahawneh, E., Al Adwan, J., Alzubi, Y. 2022. A Comparative Assessment of Regularized Regression Techniques for Modeling the Mechanical Properties of Rubberized Concrete. *The Open Civil Engineering Journal 16*(*1*), https://doi.org/10.2174/18741495-v16-e2208170

Zeleke, A.J., Palumbo, P., Tubertini, P., Miglio, R., Chiari, L. 2024. Comparison of nine machine learning regression models in predicting hospital length of stay for patients admitted to a general medicine department. *Informatics in Medicine Unlocked 47*. https://doi.org/10.1016/j.imu.2024.101499

Zhu, Q., Wen, Z., Liu, H. 2019. Microbial effects on hydraulic conductivity estimation by single-well injection tests in a petroleum-contaminated aquifer. *Journal of Hydrology 573*, 352–364. https://doi.org/10.1016/j.jhydrol.2019.03.031

Supporting Online Material