



The use of artificial neural networks in the determination of soil grain composition

Klaudia Sekuła¹ · Joanna Karłowska-Pik² · Ewa Kmiecik¹

Accepted: 16 May 2023 / Published online: 31 May 2023
© The Author(s) 2023

Abstract

The paper presents the possibility of using data mining tools — artificial neural networks — in prediction of hydrometer reading after 24 h in order to limit the duration of the test to 4 h. The authors analysed a database of 693 granulometric composition analysis results of genetically different soils with the use of radial basis function network (RBF) and multilayer perceptron (MLP). The calculations performed showed that it is possible to use MLP to shorten the test time without affecting the quality of the results. The presented accuracy of the model, in the range of 0.55–0.72, allows one to determine the content of silt and clay fractions with an accuracy of 0.49% for equivalent diameter (d_T) and 1.50% for percentage of all particles with a diameter smaller than d_T (Z_T). The results were better than that achieved using linear regression models with all predictors (REG), stepwise regression models (SREG), and classification and regression trees (CRT). Taking into account the uncertainty of hydrometric determinations, the obtained forecast values is lower than this uncertainty, therefore neural networks can be used to predict the results of this type of research.

Keywords Determination of soil grain composition · Linear regression models · Stepwise regression models · Classification and regression trees · Artificial neural networks · Radial basis function network and multilayer perceptron

1 Introduction

Data mining methods, including regression analysis and artificial neural networks (ANNs) are increasingly used to predict the physical and mechanical properties of soils and to optimize techniques of reconnaissance soil engineering properties (Yuanyou et al. 1997; Yang & Rosenbaum 2002; Boadu et al. 2013; Varghese et al. 2013).

The literature review conducted by the authors showed that often statistical tools are used to specify the relationship between individual geotechnical parameters of natural soils and rock, determined in laboratory tests (Penumadu and Zhao 1999; Lee et al. 2003; Park and Lee 2011;

Gurocak et al. 2012; Khanlari et al. 2012; Tizpa et al. 2015; Kim et al. 2021). Such correlations are also successfully used to assess the relationship between the mechanical properties of various soil mixtures and substances aimed at improving or changing them (Najjar and Basheer 1996; Debnath and Dey 2017; Dehghanbanadaki et al. 2019). The use of artificial neural networks in in-situ research is also known (Zhou and Wu 1994; Chan et al. 1995; Abu Kiefa 1998, Nejad et al. 2009; Emami and Yasrobi 2017). Statistical methods and the use of ANN are also widely used in the interpretation of geotechnical data to assess the slope stability using numerical modelling methods (Sakellariou and Ferentinou 2005; Wang et al. 2007; Mustafa et al. 2012; Lian et al. 2015; Ray et al. 2020; Li et al. 2022) or subsidence (Kanayama et al. 2014; Ghiasi and Koushki 2020).

The soil particle size analysis is performed to determine the particle size composition (ISO 2017). It allows for the determination of the percentage content of individual fractions occurring in the soil, and thus the type and name of the soil tested. Knowing the type of the tested soil is an initial laboratory test and allows to predict its properties and determine the scope of further laboratory steps to

✉ Klaudia Sekuła
klaudia.sekula@agh.edu.pl

Joanna Karłowska-Pik
joanka@mat.umk.pl

Ewa Kmiecik
ewa.kmiecik@agh.edu.pl

¹ AGH University of Krakow, Mickiewicza 30 Av.,
30-059 Kraków, Poland

² Nicolaus Copernicus University in Toruń, Toruń, Poland

specify physical and mechanical properties (Vangla and Latha 2015; Liu et al. 2020; Guo et al. 2020).

Particle size analysis can be performed by (1) the mechanical method — sieve analysis or (2) the sedimentation method — hydrometer analysis, or by using a combination of these two methods. In the case of cohesive soils, it is necessary to use the hydrometer method, which is a much more labour-intensive and time-consuming method (Barman and Choudhury 2020). In the case of sieve analysis, the preparation of the test sample is limited to drying it. Performing the hydrometer analysis requires additionally washing the dried sample, boiling the suspension and then bringing it to the ambient temperature. The further testing process for sieve analysis takes approximately 30 min. For hydrometer analysis, this time is extended to more than 24 h (Myślińska 1992, ISO 2016).

The basic objective of the article was to verify the possibility of using ANN (radial basis function network—RBF and multilayer perceptron – MLP) to predict the hydrometer reading after 24 h based on the results of measurements carried out within 4 h of mixing the suspension, which would significantly accelerate the time to conduct this type of research. An additional question was whether the measurement time could be reduced even more and how this would affect the quality of the prediction of subsequent measurements. The results obtained with the use of ANN were compared to the results obtained with the use of linear regression models with all considered predictors (REG), stepwise regression models (SREG), and classification and regression trees (CRT).

2 Materials and methods

The granulometric composition analysis were carried out at the Geotechnical and Geomechanical Research Laboratory of the Department of Hydrogeology and Engineering Geology, Faculty of Geology, Geophysics and Environmental Protection of the AGH University of Krakow. The research was conducted in 2020–2021. The database included analyses for 693 samples of natural soils of various genesis and type (from low-cohesive to very cohesive). The samples were collected at a depth between 0.5 m and 9.7 m below ground level at various points in the Śląskie, Świętokrzyskie, Małopolskie and Podkarpackie provinces. The specific density range of these soils ranged from 2.66 to 2.70, which is a typical value for cohesive soils with different content of the clay fraction. All the tested soils were mineral or low-organic soils with an organic substance value not exceeding 4.3%. The weight of the tested soil was prepared from the material previously dried to a constant weight at 105 °C and ranged from 41.25 g for very cohesive soils to 152.28 g for low-

cohesive soils with a high content of sand fraction. The average weight value was about 60 g.

The study was carried out using a combination of mechanical (sieve analysis) and sedimentation (hydrometer analysis) methods according with ISO 17892-4 (2016). Soil particles larger than 0.063 mm were subjected to sieve analysis, while smaller fractions were used to prepare the suspension and tested using the hydrometer method.

Sedimentation methods based on the fractionation of soil in water suspension are based on the Stokes law, which determines that the free-fall velocity of spherical particles is directly proportional to their diameter and specific density, and depends on the specific density and viscosity of the liquid (water) in which the particles fall and due to gravitational acceleration (Myślińska 1992).

Since Stokes law determines the falling velocity of spherical particles, and most ground particles are irregular but not spherical, the concept of equivalent diameter (d_T) is introduced.

Assuming that, at the beginning of sedimentation analysis, the carefully mixed soil suspension is homogeneous and the particles fall freely, independently of each other, we can transform the formula from Stokes's law and obtain data on the size of the falling particles. Knowing the values of the viscosity and specific density of water, the specific density of the soil skeleton, the acceleration of gravity, which are constant values for a given research, and substituting the road to time ratio for velocity, it is possible — by conducting hydrometric analysis — to determine the equivalent diameter (d_T) that has travelled the distance (HR) after the time (T). This time is specified in the ISO 14688-2 standard (ISO 2017). Hydrometer descent readings are taken after 30 s, 1, 2, 5, 15, 30 min, and after 1, 2, 4, and 24 h. The temperature of the slurry is also recorded for each reading. The percentage of particles of the calculated diameter is determined by a formula that includes measuring the density of the suspension with a hydrometer. As a result, it is obtained as a percentage of all particles with diameter smaller than d_T.

Table 1 summarizes the basic descriptive statistics of the analysed variables.

The hydrometer readings decrease over time, on average from 23.8 after 30 s to 1.0 after 24 h. The interquartile range also decreases, from 9.7 after 30 s to 3.3 after 24 h.

Comparison of skewness and its standard error indicates left-skewness of hydrometer readings after 30', 1' and 2', and right-skewness of hydrometer readings after 15', 30', 1 h, 2 h, 4 h, 24 h and temperature. Therefore, the median and interquartile range are more adequate sample characteristics.

The hydrometer reading after 24 h is strongly positively correlated (Table 2) with earlier measurements, especially with those obtained at a shorter time interval. On the one

Table 1 Descriptive statistics of analysed variables

	Variable	Mean	Std. deviation	Skewness	Skewness SE	Median	Q1 (1st quartile)	Q3 (3rd quartile)
Reading time	30''	22.49	6.84	− 0.61	0.09	23.80	17.80	27.50
	1'	21.03	6.50	− 0.67	0.09	22.40	16.50	26.00
	2'	18.72	6.19	− 0.58	0.09	20.00	14.30	23.50
	5'	15.16	6.04	− 0.06	0.09	15.10	10.95	20.00
	15'	11.54	5.93	0.39	0.09	10.60	7.00	16.30
	30'	9.56	5.62	0.56	0.09	8.40	5.30	13.90
	1 h	7.87	5.18	0.68	0.09	6.60	4.00	11.65
	2 h	6.28	4.62	0.82	0.09	5.00	2.90	9.30
	4 h	4.83	4.06	0.95	0.09	3.70	2.00	7.15
	24 h	1.83	2.98	1.28	0.09	1.00	− 0.15	3.15
Temperature [°C]		23.84	23.84	1.95	0.01	0.09	24.00	22.50

Table 2 Pearson’s correlations. * Correlation is significant at 0.05 level (2-tailed). ** Correlation is significant at 0.01 level (2-tailed)

		Reading after								Temp. [°C]
		30''	1'	2'	5'	15'	30'	1 h	2 h	4 h
Reading time	1'	0.991**								
	2'	0.939**	0.973**							
	5'	0.757**	0.828**	0.928**						
	15'	0.584**	0.670**	0.807**	0.963**					
	30'	0.520**	0.610**	0.755**	0.934**	0.994**				
	1 h	0.486**	0.575**	0.723**	0.912**	0.984**	0.996**			
	2 h	0.462**	0.551**	0.698**	0.889**	0.969**	0.986**	0.995**		
	4 h	0.450**	0.536**	0.680**	0.868**	0.951**	0.971**	0.984**	0.994**	
Temp. [°C]	0.018	0.021	0.037	0.068	0.076*	0.070	0.060	0.033	0.002	
Reading after 24 h	0.418**	0.492**	0.617**	0.778**	0.854**	0.876**	0.897**	0.920**	0.942**	−0.073

All analyses were performed using PS IMAGO PRO v. 7 software (based on the IBM SPSS Statistics v. 27 analytical engine)

hand, this can make linear methods such as linear regression very useful for predicting measurement values after 24 h from previous measurements. On the other hand, these methods may be contraindicated by the collinearity of the predictors, which is noticeable here. The hydrometer reading after 24 h is not statistically significantly correlated with temperature, which may mean that the inclusion of this predictor in the model will not translate into an improvement in its quality. The dependence of the hydrometer reading after 24 h on other features and the distributions of these features are presented in Figures S1a–j (Supplementary Material).

Several models were built for the hydrometer reading after 24 h prediction, including:

- linear regression models with all predictors (REG)

- Stepwise regression models (SREG)
- Classification and regression trees (CRT)
- Artificial neural networks — radial basis function network (RBF) and multilayer perceptron (MLP).

In all models, the predictors were the hydrometer readings after 30'', 1', 2', 5', 15', 30', 1 h, 2 h and 4 h. The usefulness of temperature as a predictor in the model was also checked.

The linear regression model for a target variable Y and predictors X_1, X_2, \dots, X_p has the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon,$$

where ε is a random error with centered normal distribution and $\beta_1, \beta_2, \dots, \beta_p$ are estimated with the least squares method. As some predictors may turn out to be statistically

insignificant, the stepwise method of selecting variables was used. In this method, in each subsequent step, a variable is added or removed according to the criterion based on the value of the F statistic. For details, see Larose and Larose (2015) and IBM SPSS (2021).

The classification and regression trees (CRT) method (Breiman et al. 1984) recursively partitions the records into subsets with similar values for the target variable. In this way, it produces graphs in which, from each decision node, starting from the initial one called the root, exactly two edges come out to the nodes on the lower level. The CRT algorithm builds the tree by conducting for each decision node an exhaustive search of all available predictors and all possible splitting values, selecting the optimal split for quantitative target variable prediction according to the least squares deviation impurity measure. The prediction of the value of the quantitative target variable for an observation that is in a given node is based on the average value of this variable for the records of the training set in that node. This means that weaker (in comparison to other methods) prediction results should be expected, since in fact the number of different possible outcomes for estimating the quantitative target variable is limited by the number of terminal nodes (leaves) in the tree. For details, see (Larose and Larose 2015; IBM SPSS 2021).

There are two models of ANN implemented in PS IMAGO PRO (IBM SPSS Statistics). The first one is the multilayer perceptron (MLP), which has an input layer, one or two hidden layers, and an output layer. For each quantitative predictor, there is one neuron in the input layer. The number of neurons in hidden layers can be automatically chosen. The output layer has one neuron for the quantitative target variable. Each neuron from a given layer is connected to all neurons from the next layer. The connections have weights assigned, which are initially numbers in the range [0; 1]. As an output from each neuron of the hidden and output layers, we obtain the value of the activation function on the linear combination of input signals and weights. The activation function for the hidden layers can be a hyperbolic tangent or sigmoid function and for the output layer additionally identity. The weights are corrected in the learning process by the backpropagation algorithm so that the error function defined as the sum of the squared errors reaches a minimum (Larose and Larose 2015; Rojas 1996).

The second model of ANN is the radial basis function (RBF) network. Compared to MLP, it has only one hidden layer in which the number of neurons depends on the number of groups that form the observations in the predictor space. Only connections between the hidden layer and the output layer have assigned weights. The weights do not require multiple corrections and are fitted by the least-squares method (IBM SPSS 2021).

Model quality was assessed using the repeated cross-validation method, which effectively increases the precision of the error estimates while still maintaining a small bias (Kuhn and Johnson 2013; James et al. 2017). The records are divided m times into k groups of similar size. For each such split, the following procedure is repeated k times. Successively, each of the k groups of records becomes a test set, and the remaining groups together are treated as a training set, on which a model is built. Then the model is checked on the test set. In this way, k measures of model quality are obtained for each of the m considered partitions. This gives together $k \times m$ measures of model quality, which are finally averaged.

As a measure of model quality, mean absolute error (MAE) and mean squared error (MSE) were considered. MAE for target variable Y is defined as:

$$\text{MAE}(Y) = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

and MSE is defined as:

$$\text{MSE}(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

where y_i denotes the observed and \hat{y}_i denotes the predicted value of the target variable Y for the i th observation ($i = 1, 2, \dots, n$, where n is the sample size) (Larose and Larose 2015).

3 Results

The final regression model built on the entire dataset is presented in Table 3. The fit of the model is very high, with the determination coefficient $R^2 = 0.921$, but, as was

Table 3 Coefficients β of the regression model built on the entire dataset and their significances

Predictor		β	Significance
Constant		−0.129	0.758
Reading after	30''	0.020	0.754
	1'	0.008	0.933
	2'	0.056	0.445
	5'	0.053	0.461
	15'	0.302	0.005
	30'	0.627	< 0.001
	1 h	0.136	0.282
	2 h	0.069	0.586
	4 h	1.382	< 0.001
Temp. [°C]		−0.027	0.118

initially supposed, most of the variables, except the readings after 15', 30' and 4 h, are statistically insignificant. The same was the case with models built in the cross-validation procedure.

Therefore, it was necessary to select variables and build models using the stepwise method with the probability of including equal to 0.05 and the probability of removing equal to 0.1. The models obtained in the cross-validation procedure contained the readings after 15', 30' and 4 h and usually one additional variable, which could be e.g. reading after 30'', 2' or 1 h. This model instability is due to the observed collinearity of the predictors.

The classification and regression tree models had a maximum depth of 5, where the minimum number of cases in the parent node was set to 10 and in the child node to 5. Most of the splits in the trees based on the reading after 4 h, and the importance of the predictors, measured as the sum of the improvements in the splits based on a given variable, was lower for the earlier measurements.

Predictors introduced into the ANN were standardized by subtracting the mean and dividing by the standard deviation (Z-score standardisation). The RBF network had softmax as the activation function in the hidden layer and identity in the output layer. The obtained networks had 4 neurons in the hidden layer, which was set automatically.

The analysis assumed a comparison of MLP models with one and two hidden layers and different activation functions in hidden (the hyperbolic tangent and the sigmoid function) and output (the hyperbolic tangent, sigmoid function and identity) layers. The entire network training procedure was carried out only on the training set. In addition, the test set was not used to determine the moment of stopping learning. The stop condition was determined by setting the number of learning epochs to 1000.

The quality of all models was evaluated using tenfold cross-validation method repeated 5 times. The average values of MAE and MSE were calculated and are presented in Table 4.

The best models of multilayer perceptrons with one and two hidden layers with sigmoid activation function in hidden layers and identity in the output layer adequately predict the hydrometer readings after 24 h, especially their positive values. For negative readings, the prediction may be slightly overestimated (Figs. 1 and 2).

MLP models with the sigmoid function in the hidden layers (one or two) and the identity function in the output layer were used to test the possibility of observation time reduction to 2 and 1 h. In each of these cases, the measurement values were predicted at further points in time. The results are presented in Tables 5 and 6.

The square root of the mean square error is 0.8264 for MLP with one hidden layer and 0.8334 for MLP with two hidden layers.

Table 4 Average MAE and MSE values obtained as a result of repeating tenfold cross-validation procedure 5 times for models of regression (REG), stepwise regression (SREG), classification and regression tree (CRT), radial basis function network (RBF), multi-layer perceptron (MLP) with one (1) or two (2) hidden layers, hiperbolic tangent (tanh) or sigmoid (sig) activation function for hidden layers and identity (id), hyperbolic tangent (tanh) or sigmoid (sig) activation function for output layer

Model	With temperature		Without temperature	
	MAE	MSE	MAE	MSE
REG	0.5654	0.7284	0.5664	0.7290
SREG	0.5648	0.7287	0.5640	0.7267
CRT	0.7286	1.1670	0.7345	1.2057
RBF	1.0859	2.2518	1.0952	2.3126
MLP 1 tanh id	0.5575	0.7056	0.5475	0.7302
MLP 1 tanh tanh	0.5551	0.7081	0.5665	0.7621
MLP 1 sig id	0.5464	0.6830	0.5473	0.7024
MLP 1 sig sig	0.5539	0.7077	0.5564	0.7469
MLP 2 tanh id	0.5537	0.7268	0.5593	0.7531
MLP 2 tanh tanh	0.5571	0.7188	0.5602	0.7516
MLP 2 sig id	0.5447	0.6946	0.5469	0.7219
MLP 2 sig sig	0.5568	0.7241	0.5554	0.7326

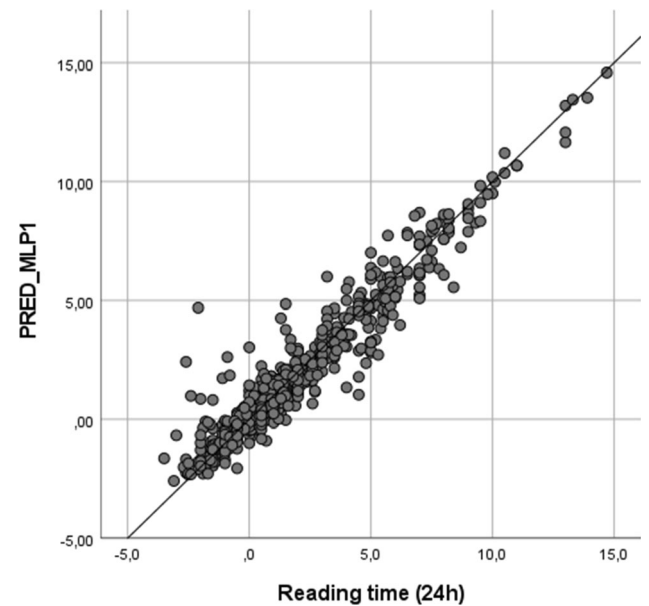


Fig. 1 Scatterplot of hydrometer reading after 24 h predicted by multilayer perceptron with one hidden layer and sigmoid activation function for the hidden layer by the true value of this reading. Predicted values obtained for one (of five) exemplary repetition of the tenfold cross-validation procedure. The line $\hat{y} = y$ is marked

4 Discussion

The labor-consumption and high cost of examining the particle size distribution using the hydrometer method has long led scientists to search for alternative methods of determining its particle size distribution. On the one hand, it is a basic research in engineering geology, however, it is

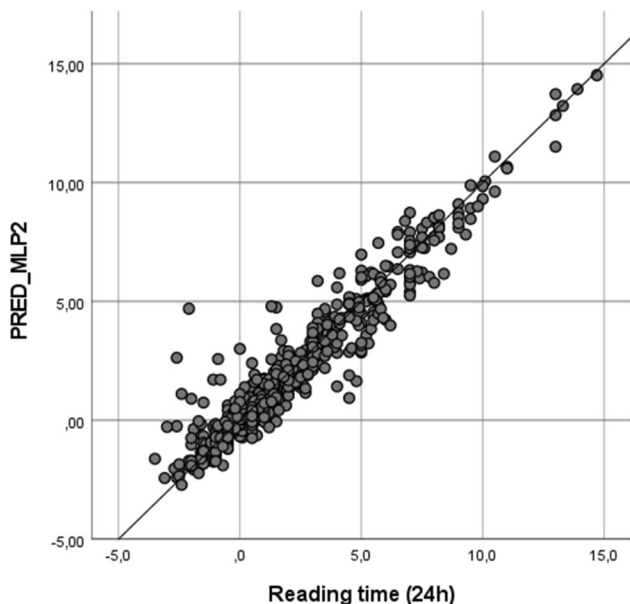


Fig. 2 Scatterplot of hydrometer reading after 24 h predicted by multilayer perceptron with two hidden layers and sigmoid activation function for the hidden layers by the true value of this reading. Predicted values obtained for one (of five) exemplary repetition of the tenfold cross-validation procedure. The line $\hat{y} = y$ is marked

Table 5 Average MAE and MSE values obtained as a result of repeating the tenfold cross-validation procedure 5 times for MLP models with one hidden layer. The observation time was successively

Observation time reduced to	MAE for reading after			MSE for reading after		
	24 h	4 h	2 h	24 h	4 h	2 h
4 h	0.5464			0.6830		
2 h	0.6322	0.2589		0.9258	0.1258	
1 h	0.7221	0.3763	0.2575	1.1695	0.2748	0.1359

Table 6 Average MAE and MSE values obtained as a result of repeating the tenfold cross-validation procedure 5 times for MLP models with two hidden layers

Observation time reduced to	MAE for reading after			MSE for reading after		
	24 h	4 h	2 h	24 h	4 h	2 h
4 h	0.5447			0.6946		
2 h	0.6293	0.2643		0.9413	0.1323	
1 h	0.7058	0.3867	0.2626	1.1384	0.2895	0.1417

The observation time was successively shortened to 4 h, 2 h and 1 h, and on this the remaining readings were predicted by the model

of key importance in the further classification and selection of more advanced laboratory tests.

Attempts to create alternative methods for determining the particle size composition were made, for example, by Barman and Choudhury (2020), who presented classification of soil images using multi SVM and linear kernel function. However, they emphasize the adequacy of the use of this method solely for the purpose of determining the texture characteristics of soils for the purposes of agriculture. The accuracy of the system has not been verified in the case of determining the content of the exact sizes of individual soil fractions which is of key importance in the geotechnical classification. Ghasemy et al. (2019) proposed a mathematical approach based on comparing the results of the combination of sedimentation and spectrophotometric methods. However, laboratory tests were performed on only 17 samples and despite the confirmation of the initial assumptions, the accuracy of the results was not determined. This makes it impossible to compare the results of this experimental method for all types of soil.

Owji et al. (2014) in their publication showed that in the case of hydrometer readings using the Bouyoucos method, it is possible to shorten the reading time to 2 h, but only to determine the overall texture of the soil. However, such a procedure is not sufficiently precise. They also showed that each subsequent hydrometer reading significantly influences the determination of the content of the finer fractions. And it is their number that is of key importance in the final classification of cohesive soils, which is very extensive and requires the specification of the content of clay and silt fractions with high accuracy. Adiku et al. 2005 indicated

shortened to 4 h, 2 h and 1 h, and on this the remaining readings were predicted by the model

Table 7 The mean values of the differences between the real dT and ZT values and the dT and ZT values calculated from the predicted R values based on MLP with one hidden layer (MLP 1 sig id) and MLP

with two hidden layers (MLP 2 sig id) with sigmoid activation function in hidden layers and identity in the output layer

Parameter	Number of analysed samples	MLP 1 sig id	MLP 2 sig id
The mean value of the difference between the real dT value and the value calculated from the predicted R readings	693	0.49	0.49
The mean value of the difference between the real ZT value and the value calculated from the predicted R readings for the entire data set	693	1.50	1.49
The mean value of the difference between the real ZT value and the value calculated from the predicted R readings for $R_{\text{real}} \leq 5$	594	1.38	1.37
The mean value of the difference between the real ZT value and the calculated value from the predicted R readings for $R_{\text{real}} > 5$	99	2.24	2.15

that the hydrometer readings at any time can be predicted from the exponential equation provided that the reading after 4.5 min ($R_{4.5}$) and the experimentally determined exponent B are known. The accuracy between the calculated and measured R values was determined using the equations defined by them as $R_2 = 0.96$. However, this method is not universal for all types of soils, but only for those that are similar in type and genesis.

Used by Fragomeni et al. (2021) multiple regression analysis and stepwise regression analysis in the evaluation of the relationship between different geotechnical parameters showed the possibility of developing predictive models, the effectiveness and reliability of which are better than others. They also indicated that their use saves time and money in laboratory research. On the other hand, Gołębiewska and Hyb (2008) formulated conclusions regarding the uncertainty of the parameters determined in the hydrometer analysis, i.e. the equivalent diameter (dT) and their content (ZT). They determined the mean uncertainty of dT equal to 3% ($dT \pm 3\% dT$) and the uncertainty of the particle content (ZT) equal to 8% of the particle size at R hydrometer readings below 5 ($ZT \pm 8\% ZT$) and at larger readings equal to 3% particle size ($ZT \pm 3\% ZT$).

The determination of the uncertainty of the dT and ZT values for the models made in this research was carried out by comparing the actual value of the equivalent grain diameter (dT_{real}), calculated on the basis of the actual R reading (R_{real}), to the dT values calculated on the basis of the R value provided in the models. The predicted R value obtained as a result of the fivefold validation was averaged for MLP with the sigmoid function in the hidden layers (one – R_ONE or two – R_TWO) and the identity function in the output layer. Then, on their basis, dT_{ONE} , dT_{TWO} , ZT_{ONE} and ZT_{TWO} were calculated.

Subsequently, the differences between the real dT and ZT values and the calculated value from the predicted R readings were determined. The results are summarized in Table 7.

The presented results show that both the dT and ZT values, calculated on the basis of the predicted values of the R, are within the acceptable limits for the uncertainty of these determinations. In the case of ZT, a higher error is noticeable for $R_{\text{real}} > 5$ readings. This is due to the much smaller number of samples whose reading after 24 h exceeded this value. A similar value of the mean error for the entire data set to the error value for the R readings ≤ 5 shows that the solution presented in the article can be used successfully for the entire data set without differentiating the samples due to the value of the last R reading.

The machine learning methods used in the analyses in this article (linear regression, CRT and MLP) treat the hydrometer readings as separate predictors, they do not take into account in any way the fact that they were read in a specific order, at moments of time separated by a known number of minutes. Probably the problem of predicting hydrometer readings could also be analysed as time series forecasting problem, for which dedicated more advanced techniques can be used (eg. recurrent neural networks). This may be the subject of further research. However, it should be borne in mind that the results may not be satisfactory. This is due to the fact that the studied time series are short and consist of observations at only a few time points. Moreover, the analysed time series (hydrometer readings) are monotonic (non-increasing) what simplifies the situation and means that less advanced methods may be sufficient.

5 Conclusions

Optimization of the research process can be achieved through the construction of new equipment or the improvement of existing equipment or research method. However, the process is not simple and fast. They are also often unfavorable or not financially optimal solutions. The use of statistical tools, including neural networks, is a much

simpler, faster and cheaper solution. It only requires a sufficient amount of data and basic statistical software.

The methodology presented in the article presents the possibility of using neural networks in the prediction of hydrometer readings after 24 h, which allows for significant shortening of the test and optimization of laboratory procedures without compromising the credibility of the obtained results. The calculations performed also showed the possibility of predicting the hydrometer readings after 1 h, 2 h and 4 h. In the case of readings for these times, the accuracy is lower, but it can still be used to determine the grain size composition for soils with less differentiation and a lower content of clay and silt fractions.

Considering the uncertainty of hydrometric determinations, the obtained forecast value is lower than this uncertainty, therefore neural networks can be used to predict the results of this type of research. However, the condition for the laboratory to use neural networks to predict readings is to collect a sufficiently large database of full hydrometric test readings for soils that may differ in type and origin, but occur in a defined geographical area. It is also recommended to periodically update and calibrate the results by performing control tests.

Author contributions Conceptualization, K.S.; methodology, K.S. and J.K.P.; data curation, K.S.; validation, K.S., J.K.P. and E.K.; formal analysis, K.S.; investigation, K.S.; writing—original draft preparation, K.S. and J.K.P.; writing—review and editing, K.S., J.K.P. and E.K.; supervision, E.K. All authors have read and agreed to the published version of the manuscript.

Funding The study was partially financed by AGH-UST 16.16.140.315/10.

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00477-023-02480-7>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright

holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abu Kiefa MA (1998) General regression neural networks for driven piles in cohesionless soils. *J Geotech Geoenviron Eng.* [https://doi.org/10.1061/\(ASCE\)1090-0241\(1998\)124:12\(1177\)](https://doi.org/10.1061/(ASCE)1090-0241(1998)124:12(1177))
- Adiku SGK, Osei G, Adjadeh TA, Dowuona GN (2005) Simplifying the analysis of soil particle sizes I. Test of the Sur and Kukal's modified hydrometer method. *Commun Soil Sci Plant Anal.* <https://doi.org/10.1081/LCSS-200026828>
- Barman U, Choudhury RD (2020) Soil texture classification using multi class support vector machine. *Inf Process Agric.* <https://doi.org/10.1016/j.inpa.2019.08.001>
- Boadu FK, Owusu-Nimo F, Achampong F, Ampadu SI (2013) Artificial neural network and statistical models for predicting the basic geotechnical properties of soils from electrical measurements. *Near Surf Geophys* 11:599–612
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall/CRC, New York
- Chan WT, Chow YK, Liu LF (1995) Neural network: an alternative to pile driving formulas. *Comput Geotech.* [https://doi.org/10.1016/0266-352X\(95\)93866-H](https://doi.org/10.1016/0266-352X(95)93866-H)
- Debnath P, Dey AK (2017) Prediction of laboratory peak shear stress along the cohesive soil-geosynthetic interface using artificial neural network. *Geotech Geol Eng.* <https://doi.org/10.1007/s10706-016-0119-2>
- Dehghanbanadaki A, Sotoudeh MA, Golpazir I (2019) Prediction of geotechnical properties of treated fibrous peat by artificial neural networks. *Bull Eng Geol Environ.* <https://doi.org/10.1007/s10064-017-1213-2>
- Emami M, Yasrobi SS (2017) Modeling and interpretation of pressuremeter test results with artificial neural networks. *Geotech Geol Eng.* <https://doi.org/10.1007/s10706-013-9720-9>
- Fragomeni C, Hedayat A, Asce AM, Navidi W, Kuhn E, Thomas D, Perkin, M (2021) Development of prediction models for resilient modulus of soils. Rocky mountain geo-conference 2021
- Ghasemy A, Rahimi E, Malekzadeh A (2019) Introduction of a new method for determining the particle-size distribution of fine-grained soils. *Measurement.* <https://doi.org/10.1016/j.measurement.2018.09.041>
- Ghiasi V, Koushki M (2020) Numerical and artificial neural network analyses of ground surface settlement of tunnel in saturated soil. *SN Appl Sci.* <https://doi.org/10.1007/s42452-020-2742-z>
- Gołębiewska A, Hyb W (2008) Ocena niepewności wyników pomiarów w analizie areometrycznej gruntu. *Geoinżynieria* 4:30–35 ([in Polish])
- Guo Z, Lai J, Jin J, Zhou J, Zhao K, Sun Z (2020) Effect of particle size and grain composition on two-dimensional infiltration process of weathered crust elution-deposited rare earth ores. *T Nonferrous Metal Soc.* [https://doi.org/10.1016/S1003-6326\(20\)65327-4](https://doi.org/10.1016/S1003-6326(20)65327-4)
- Gurocak Z, Solanki P, Alemdag S, Zaman MM (2012) New considerations for empirical estimation of tensile strength of rocks. *Eng Geo.* <https://doi.org/10.1016/j.enggeo.2012.06.005>
- IBM SPSS statistics algorithms. Available on-line: https://www.ibm.com/docs/en/SSLVMB_27.0.0/pdf/en/IBM_SPSS_Statistics_Algorithms.pdf. Accessed on 4th Oct 2021
- ISO 14688-2 (2017) Geotechnical investigation and testing — Identification and classification of soil — Part 1: identification and classification of soil. Principles for a classification

- ISO 17892-4 (2016) Geotechnical investigation and testing — Laboratory testing of soil — Part 4: determination of particle size distribution
- James G, Witten D, Hastie T, Tibshirani R (2017) An introduction to statistical learning with applications in R. Springer, New York
- Kanayama M, Roh A, Paassen LA (2014) Using and improving neural network models for ground settlement prediction. *Geotech Geol Eng*. <https://doi.org/10.1007/s10706-014-9745-8>
- Khanlari GR, Heidari M, Momeni AA, Abdilor Y (2012) Prediction of shear strength parameters of soils using artificial neural networks and multivariate regression methods. *Eng Geol*. <https://doi.org/10.1016/j.enggeo.2011.12.006>
- Kim Y, Satyanaga A, Rahardjo H, Park H, Lun Sham AW (2021) Estimation of effective cohesion using artificial neural networks based on index soil properties: a Singapore case. *Eng Geol*. <https://doi.org/10.1016/j.enggeo.2021.106163>
- Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York
- Larose DT, Larose CD (2015) Data mining and predictive analytics, 2nd edn. Wiley, New Jersey
- Lee SJ, Lee SR, Kim YS (2003) An approach to estimate unsaturated shear strength using artificial neural network and hyperbolic formulation. *Comput Geotech*. [https://doi.org/10.1016/S0266-352X\(03\)00058-2](https://doi.org/10.1016/S0266-352X(03)00058-2)
- Li Y, Rahardjo H, Satyanaga A, Rangarajan S, Tsen-Tieng Lee D (2022) Soil database development with the application of machine learning methods in soil properties prediction. *Eng Geol*. <https://doi.org/10.1016/j.enggeo.2022.106769>
- Lian C, Zeng Z, Yao W, Tang H (2015) Multiple neural networks switched prediction for landslide displacement. *Eng Geol*. <https://doi.org/10.1016/j.enggeo.2014.11.014>
- Liu X, Zou D, Liu J, Zhou C, Zheng B (2020) Experimental study to evaluate the effect of particle size on the small strain shear modulus of coarse-grained soils. *Measurement*. <https://doi.org/10.1016/j.measurement.2020.107954>
- Mustafa MR, Rezaur RB, Rahardjo H, Isa MH (2012) Prediction of pore-water pressure using radial basis function neural network. *Eng Geol*. <https://doi.org/10.1016/j.enggeo.2012.02.008>
- Myślińska E (1992) Laboratoryjne badania gruntów. Wydawnictwo Naukowe PWN, Warszawa [in Polish]
- Najjar YM, Basheer IA (1996) Utilizing computational neural networks for evaluating the permeability of compacted clay liners. *Geol Eng, Geotech*. <https://doi.org/10.1007/BF00452947>
- Owji A, Esfandiarpour Boroujeni I, Kamali A, Hosseiniard SJ, Bodaghabadi MB (2014) The effects of hydrometer reading times on the spatial variability of soil textures in Southeast Iran. *Arab J Geosci*. <https://doi.org/10.1007/s12517-012-0786-0>
- Park HI, Lee SR (2011) Evaluation of the compression index of soils using an artificial neural network. *Comput Geotech*. <https://doi.org/10.1016/j.compgeo.2011.02.011>
- Penumadu D, Zhao R (1999) Triaxial compression behavior of sand and gravel using artificial neural networks (ANN). *Comput Geotech*. [https://doi.org/10.1016/S0266-352X\(99\)00002-6](https://doi.org/10.1016/S0266-352X(99)00002-6)
- Pooya Nejad F, Jaksza MB, Kakhi M, McCabe BA (2009) Prediction of pile settlement using artificial neural networks based on standard penetration test data. *Comput Geotech*. <https://doi.org/10.1016/j.compgeo.2009.04.003>
- Ray A, Kumar V, Kumar A, Rai R, Khandelwal M, Singh TN (2020) Stability prediction of Himalayan residual soil slope using artificial neural network. *Nat Hazards*. <https://doi.org/10.1007/s11069-020-04141-2>
- Rojas R (1996) Neural networks. Springer, Berlin, A systematic Introduction
- Sakellariou MG, Ferentinou MD (2005) A study of slope stability prediction using neural networks. *Geotech Geol Eng*. <https://doi.org/10.1007/s10706-004-8680-5>
- Tizpa P, Jamshidi Chenari R, Karimpour Fard M, Lemos Machado S (2015) ANN prediction of some geotechnical properties of soil from their index parameters. *Arab J Geosci*. <https://doi.org/10.1007/s12517-014-1304-3>
- Vangla P, Latha GM (2015) Influence of particle size on the friction and interfacial shear strength of sands of similar morphology. *Int J Geosynth Ground Eng*. <https://doi.org/10.1007/s40891-014-0008-9>
- Varghese VK, Babu SS, Bijukumar R, Cyrus S, Abraham BM (2013) Artificial neural networks: a solution to the ambiguity in prediction of engineering properties of fine-grained soils. *Geotech Geol Eng*. <https://doi.org/10.1007/s10706-013-9643-5>
- Wang Z, Li Y, Shen RF (2007) Correction of soil parameters in calculation of embankment settlement using a BP network back-analysis model. *Eng Geol*. <https://doi.org/10.1016/j.enggeo.2007.01.007>
- Yang Y, Rosenbaum MS (2002) The artificial neural network as a tool for assessing geotechnical properties. *Geotech Geol Eng*. <https://doi.org/10.1023/A:1015066903985>
- Yuanyou X, Yanming X, Ruigeng Z (1997) An engineering geology evaluation method based on an artificial neural network and its application. *Eng Geol*. [https://doi.org/10.1016/S0013-7952\(97\)00015-X](https://doi.org/10.1016/S0013-7952(97)00015-X)
- Zhou Y, Wu X (1994) Use of neural networks in the analysis and interpretation of site investigation data. *Comput Geotech* 16(2):105–122. [https://doi.org/10.1016/0266-352X\(94\)90017-5](https://doi.org/10.1016/0266-352X(94)90017-5)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.