



Machine learning-based estimation of soil's true air-entry value from GSD curves



Mohammad Sadegh Es-haghi^{a,b}, Mohammad Rezania^{c,*}, Meghdad Bagheri^d

^a Escuela Técnica Superior de Ingenieros Navales (ETSIN), Universidad Politécnica de Madrid (UPM), Av. de la Memoria, 4, 28040 Madrid, Spain

^b Centre Internacional de Mètodes Numèrics a l'Enginyeria (CIMNE), Edifici C1, Campus Norte, UPC, Gran Capitán s/n, 08034 Barcelona, Spain

^c School of Engineering, University of Warwick, Coventry CV4 7AL, UK

^d School of Energy, Construction and Environment, Coventry University, Coventry CV1 5FB, UK

ARTICLE INFO

Article history:

Received 16 March 2022

Revised 26 May 2022

Accepted 22 June 2022

Available online 27 June 2022

Keywords:

Air-entry value

Soil water retention

Grain size distribution

Machine learning

ABSTRACT

The application of machine learning (ML) methods has proven to be promising in dealing with a wide range of geotechnical engineering problems in recent years. ML methods have already been used for the prediction of soil water retention curves (SWRC) and estimation of air-entry values (AEV). However, the reported works in the literature are generally based on limited data and conventional, less accurate approaches for AEV estimation. In this paper, a large database, known as UNSaturated SOil hydraulic DAtabase (UNSODA), is studied and the conventional and true AEVs of 790 soil samples are estimated based on determination methods reported in the literature. A ML approach is then employed for the development of a predictive model for the estimation of true AEV from water content-based SWRCs of a wide range of soil types taking into account the impact of bulk density and grain size distribution parameters. The obtained results reveal an enhanced accuracy in AEV determination, featuring R^2 values of 0.964, 0.901 and 0.851 for training, validation, and testing data, respectively, which confirm the high-level performance of the developed ML model. Based on the results of a sensitivity analysis, the particle sizes of 50 and 250 μm are found to have the highest impact on the AEV estimation.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of International Association for Gondwana Research. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In unsaturated soil mechanics, the relationship between the amount of water held by the soil and the pore-water tension (suction) is generally presented in the form of water content (w) versus suction (s), or degree of saturation (S_r) versus suction in a semi-logarithmic graph known as soil water retention curve (SWRC). Generally, three distinct parts are recognized on a SWRC (Fig. 1). Initially, with the soil undergoing drying, the curve remains a relatively flat line until the air-entry value (AEV); during this stage, with an increase of suction, the pores remain saturated ($S_r = 100\%$) with pore-water being under tension (saturated state or boundary effect zone). As suction exceeds the AEV, air breaks into the larger pores, and a continuous network of air-filled and water-filled pores is formed. During this stage, the degree of saturation (or water content) is progressively decreased as more water evaporates or leaves the pores (partially saturated state or transition effect zone). Finally, with an increase in suction, a state is

reached where water only remains at the particle contacts and is no longer continuous with the pore space. At this stage, the SWRC almost flattens meaning that much less water will be expelled with an increase in suction, and the soil dries without significant volume changes (residual state). The two inflection points, namely AEV and residual suction (s_{rs}), are considered as fundamental parameters for the determination of soils' water retention properties, with the former being a key input parameter in several unsaturated constitutive models (e.g., Alonso et al. 1990; Russell and Khalili 2006), hence, highlighting the importance of its accurate estimation from SWRCs.

On a degree of saturation-based SWRC, the AEV is graphically determined as the intersection point of a horizontal tangent line drawn to the saturated portion, with a tangent line drawn to the transition portion of the curve (see Fig. 1). Pasha et al. (2016) showed that using the same method to derive AEV from water content-based SWRCs can produce highly erroneous results. They proposed a simple method for estimation of true AEV from water content-based SWRC. This method was based on plotting the data on both semi-log and log-log scales on the same graph. In this way, the saturated part of the curve is identified by a linear or

* Corresponding author.

E-mail address: m.rezania@warwick.ac.uk (M. Rezania).

Nomenclature

P	particle size	GSD	grain size distribution
\hat{p}_c	preconsolidation pressure	GP	genetic programming
R^2	coefficient of determination	HCT	high-capacity tensiometer
s	soil suction	MEP	multi-expression programming
s_{ae}	suction at air-entry	ML	machine learning
s_{rs}	residual soil suction	MLP	multilayer perceptron
S_r	degree of saturation	RF	random forest
w	gravimetric water content	RMSE	root mean square error
θ	volumetric water content	SVM	support vector machine
AEV	air-entry value	SWRC	soil water retention curve
AI	artificial intelligence	UNSODA	UNsaturated SOil hydraulic DATabase
ANN	artificial neural network	USDA	United States Department of Agriculture
EPR	evolutionary polynomial regression		

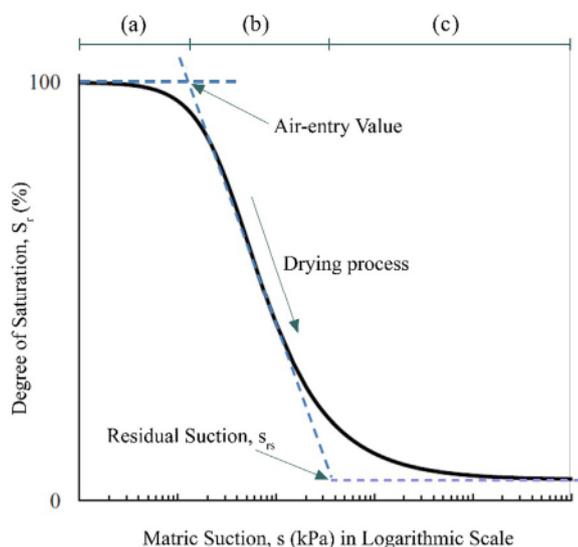


Fig. 1. Typical SWRC in $S_r - s$ plane: (a) boundary effect zone; (b) transition zone; (c) residual zone.

bilinear behavior in a semi-log plot, and the unsaturated part of the curve is represented by a straight line (linear behavior) on the log-log plot. Therefore, by drawing these two complementary curves on a single graph, the AEV as the boundary between saturated and transition (unsaturated) zones can be readily identified. This graphical technique can be used for the evaluation of true AEV from SWRC data where the information on the volume change of the sample during testing is not available.

SWRCs are generally developed based on experimental methods such as axis translation (Bagheri et al., 2019a), negative water column (Pagano et al. 2016) and pressure plate (Tarantino et al. 2011). Such methods are generally time-consuming and expensive, and in most cases produce discontinuous measurements, bringing difficulties in the accurate determination of the AEV. For such measurements, the method proposed by Pasha et al. (2016) appears promising. Bagheri (2018), Bagheri et al. (2018), and Bagheri and Rezania (2022) later showed that direct measurement of soil suction changes using high-capacity tensiometers (HCT) can rectify the need for such approximation methods, as an accurate estimation of AEV can be readily obtained from continuous measurements of suction variations with water content. However, this method is also limited to the maximum capacity of HCTs which is typically in the range of 1.5–2.0 MPa. Other methods including estimation of SWRC based on the soils' grain size distribution

(GSD) curve (Alves et al. 2020; Zhai et al. 2020), statistical methods (Saxton et al. 1986; Chiu et al. 2012), and artificial intelligence (AI)-based methods (Schaap and Leij 1998) are also available. However, it is apparent that accurate estimation of AEV is profoundly subordinate to the accuracy of predicted SWRCs. Furthermore, the reported works on statistical and AI methods generally utilized databases that were not large enough to include a variety of soil types. It is therefore imperative to consider a relatively large database to directly estimate true AEV from common soil parameters.

Recently, machine learning (ML) approaches have proven to be promising in solving nonlinear and complicated problems using large databases (Rezania 2008; Javadi and Rezania 2009a; Zhang et al. 2020; Zhang et al. 2021; Wang et al., 2020b; Zhang et al. 2022; Wang et al., 2020c). In geotechnical engineering, the ML algorithms have been favorably employed for predicting various phenomena such as the settlement of shallow foundations on cohesionless soils (Rezania and Javadi 2007), thermo-hydro-mechanical behavior of hydrate reservoirs (Zhou et al. 2020), non-stationary and non-Gaussian geotechnical properties (Shi and Wang 2021), soil constitutive modeling (Javadi and Rezania 2009b; Rezania and Ma 2019) and suction distribution in shallow soil layers (Cheng et al. 2020). In these studies, various ML algorithms including support vector machine (SVM), multi-expression programming (MEP), genetic programming (GP), evolutionary polynomial regression (EPR), and random forest (RF) have been used. The effectiveness of ML methods has led to their employment in the estimation of SWRCs (Jain et al. 2004; Moreira de Melo and Pedrollo 2015). A few works have been also reported in the literature on the estimation of the SWRC from GSD curves employing ML methods (D'Emilio et al. 2018; Amanabadi et al. 2019; Li and Vanapalli, 2022). However, estimation of the AEV by ML algorithms has been rarely studied. Recently, Wang et al. (2020a) utilized ML algorithms for the prediction of AEV of compacted soils based on some physical properties. The authors have used the parameters of sand content, fines content, plasticity index, initial water content and initial void ratio as input variables. In addition, they consider the conventional and less accurate method of estimating AEV from water content-based SWRC which can be considered as the shortcoming of their study. With these explanations, a thorough understanding of ML algorithms in the prediction of the true AEV is worth investigating.

In this paper, the UNsaturated SOil hydraulic DATabase (UNSODA) (Leij et al. 1996) has been thoroughly reviewed and for a considerable number of soil samples, the true AEVs are estimated from the water content-based SWRCs using the method proposed by Pasha et al. (2016) and compared to those obtained from the conventional method for AEV determination. Furthermore, the influence of physical soil parameters (e.g., bulk density,

grain size, etc.) on AEV is investigated. Finally, a neural network model is employed to estimate the true AEVs. Sensitivity analysis is also performed in order to determine the consideration of which parameters are more significant for the AEV prediction.

2. UNSODA

UNSODA is a collection of data for unsaturated hydraulic soil properties in Microsoft Access-97 format. The data for 790 soil samples have been categorized in 36 tables, arranging them in rational groups with relevant information. A 4-digit code has been allocated to each soil sample. Fig. 2 shows the frequency histogram of soil classification of the 790 samples according to USDA (United States Department of Agriculture) system and the summary of their geographical distribution represented in UNSODA. Most of the samples of the database are from Europe and North America. The majority of the soil samples are coarse-textured although there is an acceptable number of fine soil samples.

UNSODA provides a wide range of soil properties as shown in Table 1. It must be noted here that bulk density, particle density, porosity, saturated conductivity, and saturated water content are the only parameters required for the estimation of true AEV in the present study. In addition, UNSODA provides appropriate information about the grain size distribution of samples; however, there is no uniform set of particle sizes in the UNSODA, and the reported values of particle fraction are presented for different sizes in different samples due to the differences in experimental methods. For instance, in soil sample 1090, particle fractions for the particle sizes of 2, 50, 125, 250, 500, 1000, and 2000 μm have been determined; however, in the soil sample 1087 only the particle size of 2, 50, and 2000 μm have been considered.

In this study, the particle size fractions of 2, 20, 50, 250, 500, 1000, and 2000 μm, which have the most frequency in the database are used. Fig. 3 presents the frequency of particle sizes in the UNSODA database.

Table 2 presents the number of hydraulic curves and the number of data pairs partitioned into drying and wetting parts, as well as curves obtained from laboratory or field measurements. In this study, the water retention curves (a total of 902 curves) which have the most data are used. From the 902 water retention (h-θ) curves, the drying curves, with a total number of 867 curves from both laboratory and field measurements, are considered for the analysis.

Fig. 4 presents the scatter plot of the main drying branch of SWRCs reported in the UNSODA database for both field and lab

measurements. In this Figure, the range of changes and how suction and water content data are distributed for the lab and field data are well illustrated. As it turns out, in the lab-based works, the results involve a wider range of suction and water content values.

3. True air entry value

Although many researchers recognize the first break in the drying branch of the water content-based SWRC as evidence of air-entry point (Fredlund and Xing 1994; Zhai and Rahardjo 2012; Bagheri et al., 2019b, Rezania et al., 2020), this may not be true for some deformable soils, depending on factors such as the mechanical characteristics of the soil, pore-size distribution index, and the stress history of the soil. In the graphical method presented by Pasha et al. (2016), estimation of the true AEV was carried out based on the stress history of the soil, and consideration of three stress states, namely normally consolidated, overconsolidated with $s_{ae} > p'_c$, and overconsolidated with $s_{ae} < p'_c$, where s_{ae} is the suction at the air-entry and p'_c is the preconsolidation pressure of the soil.

From the 867 SWRCs in the UNSODA database, 221 curves could not be used for the analysis for various reasons, such as an insufficient number of data points or soil samples not entering the transition zone. Therefore, only SWRCs that would allow for appropriate graphical estimation of the AEV are selected and considered in his study. Of the remaining 644 curves, 212 curves are related to the normally consolidated state, 182 curves are related to the overconsolidated state with $s_{ae} > p'_c$, and 250 curves are related to the overconsolidated state with $s_{ae} < p'_c$. As shown in Fig. 5, for all these 644 curves, the AEVs are obtained following the graphical approach proposed by Pasha et al. (2016) (denoted as true AEV) and also according to the conventional method of finding the intersection of the tangent lines to the boundary effect zone and the transition zone (denoted as AEV). In this figure, the horizontal axis represents the codes allocated to the samples, which can be traced in the UNSODA database. With some examples, Fig. 6 illustrates the differences between the evaluated AEVs from the two methods. In addition, Fig. 7 presents the frequency of variations between the true AEVs and the conventional AEVs which for the majority of cases appear to be in the range of 30 to 300 kPa.

4. Preparing the database

The recorded data in the UNSODA database is presented in a way that is not convenient for training the ML model. On the other

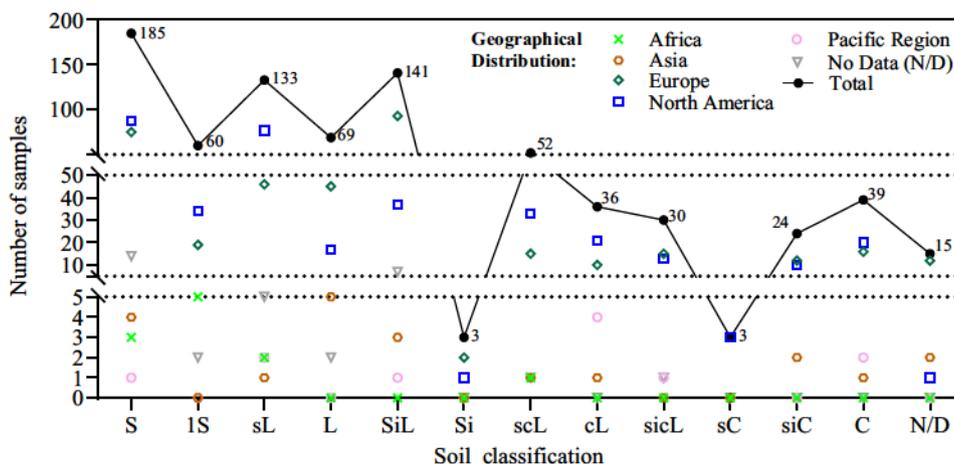


Fig. 2. Geographical and textural distribution according to the USDA classification of soils in UNSODA (S: Sand, IS: loamy Sand, sL: sandy Loam, L: Loam, SiL: silty Loam, Si: Silt, scL: sandy clay Loam, cL: clay Loam, sicL: silty clay Loam, sC: sandy Clay, siC: silty Clay, C: Clay, N/D: Not Determined).

Table 1
Soil properties provided in UNSODA database.

Parameter	Definition	Available Values	Missing Values	Considered in this study?
bulk density	Bulk density as a mass of solids per bulk volume	762	28	Yes
particle density	Mass of solids per volume of solids	339	391	Yes
porosity	Volume of voids per bulk volume	270	420	Yes
OM content	Organic Matter Content. The mass of organic matter content as a percentage of the total solid mass.	388	402	No
k_{sat}	Saturated Conductivity. The measured saturated hydraulic conductivity	429	361	Yes
θ_{sat}	Saturated Water Content. The experimental water content of a water-saturated sample	305	485	Yes
CEC	Cation Exchange Capacity, in cmol of charge per kg of dry soil (i.e., meq/100 g soil)	150	640	No
pH	Measured soil pH	300	490	No
electrolyte level	The approximate total solute concentration of the soil solution during the experiments	26	764	No
SAR	Sodium Adsorption Ratio	80	710	No
ESP	Exchangeable Sodium Percentage	19	771	No
EC	Electrical conductivity of the saturation extract	62	728	No
free Fe Al oxide	The mass fraction of the Fe and Al oxides as a percentage of the total solid phase	14	776	No

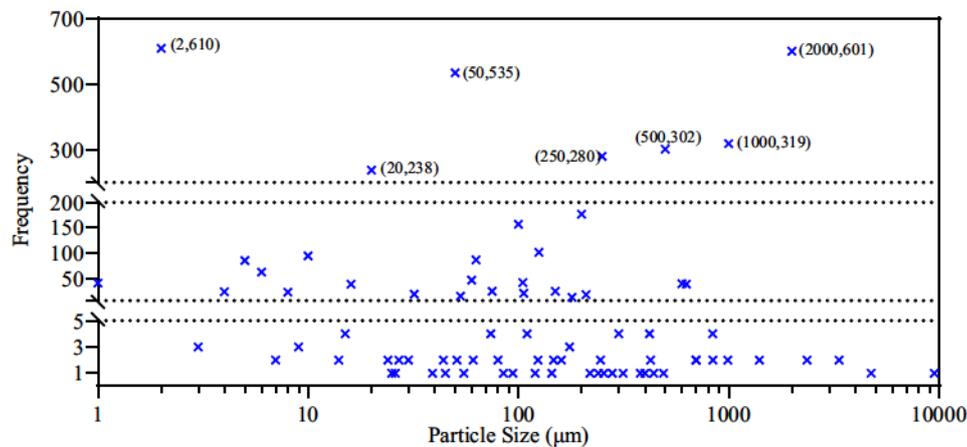


Fig. 3. Frequency of the particle size in the UNSODA database.

Table 2
Summary of the number of hydraulic curves.

Hydraulic Curves		Number of Curves		Total Number of Data Pairs		Average Number of Data Pairs in each Curve	
		Field	Lab	Field	Lab	Field	Lab
Water Retention (h- θ)	Drying	137	730	2621	8066	19.1	11.0
	Wetting	2	33	8	528	4.0	16.0
	Total	902		11,223		12.4	
Hydr. Conductivity (h-K)	Drying	133	730	2826	6187	21.2	8.5
	Wetting	0	8	0	71	-	8.9
	Total	871		9084		10.4	
Hydr. Conductivity (θ -K)	Drying	294	293	5391	5177	18.3	17.7
	Wetting	0	20	0	216	-	10.8
	Total	607		10,784		17.8	
Soil Water Diffusivity (θ -D)	Drying	56	92	1282	1456	22.9	15.8
	Wetting	0	2	0	13	-	6.5
	Total	150		2751		18.3	

hand, as shown in Table 1, in this database some of the parameters are not reported for all samples. For instance, the data points of the GSD curves are presented in different particle sizes (see Fig. 3). Therefore, at first, to extract suitable information from the database and arrange it in the way that is useful to train the neural networks, a sorting code is developed in MATLAB software. The developed code also allows for extrapolation of the missing data and formation of the complete database required for the ML.

As shown in Fig. 3, the most common particle size values in UNSODA are $P_{2\mu}$, $P_{20\mu}$, $P_{50\mu}$, $P_{250\mu}$, $P_{500\mu}$, $P_{1000\mu}$, and $P_{2000\mu}$. As a result, the fraction corresponding to these values are selected from the GSD curve as part of the ML inputs, and for samples that do not have these particle sizes, the corresponding values are estimated using the method proposed in Vaz et al. (2020). To overcome the common limitation of the GSD curves, which is the lack of standardization for granulometric fractions collected from various soil

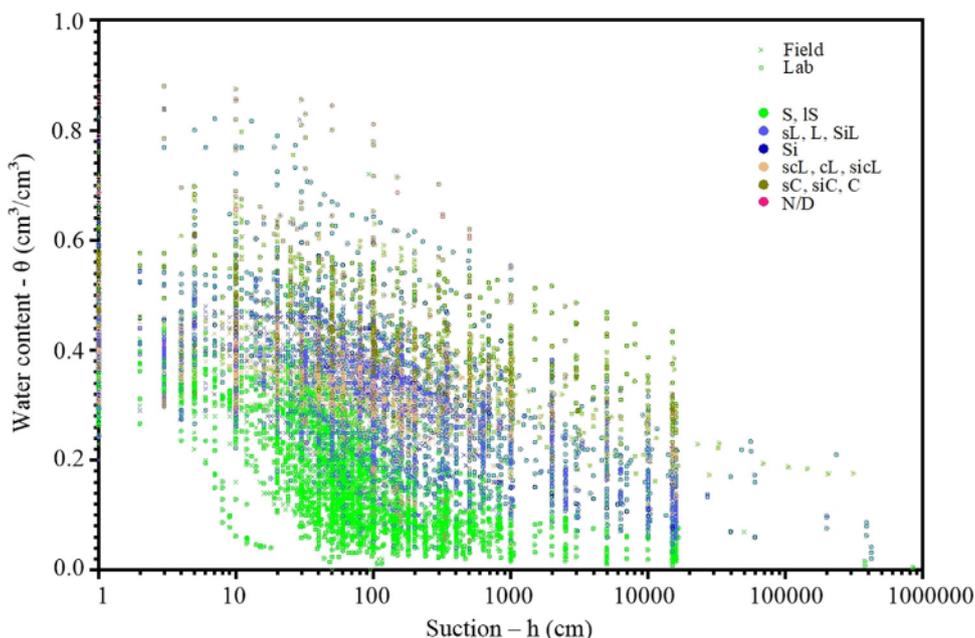


Fig. 4. Main drying branch of SWRCs in UNSODA database (S: Sand, IS: loamy Sand, sL: sandy Loam, L: Loam, SiL: silty Loam, Si: Silt, sCL: sandy clay Loam, cL: clay Loam, sicL: silty clay Loam, sC: sandy Clay, siC: silty Clay, C: Clay, N/D: Not Determined).

analysis methods, Vaz et al. (2020) evaluated and compared the performance of several GSD equations and showed that the three-parameter Eq. (1) has an acceptable performance with root mean square error (RMSE) of 0.463, 0.205, and 0.013, respectively for sand, silt, and clay. Therefore, this equation is adopted in this study:

$$F(d) = \left[1 + \left(\frac{a}{d} \right)^b \right]^{-c} \tag{1}$$

In the above equation, *a*, *b*, and *c* are equation parameters, *d* is the particle size and *F(d)* is the fraction corresponding to the particle size of *d*. Thus, for each sample, the parameters of Eq. (1) are extracted and the fraction values in the required particle size (if not present) are calculated. For instance, in Fig. 8, the fitted curve for sample 3340 in the database is presented and the values of *P*_{20μ} and *P*_{50μ} for this sample are estimated. This process is repeated for all samples in order to complete the database.

The detailed statistics of all variables in the database, with the values of minimum, maximum, median, mean, and standard deviation are summarized in Table 3 and their descriptive frequency histograms are shown in Fig. 9. Bulk density and porosity have the highest frequency at 1.4–1.6 and 0.4–0.5 respectively and follow an approximately normal distribution (Fig. 9a and b). Regarding mass fractions (Fig. 9c–j), it is observed that the highest frequency of *P*_{2μ} is between 0.0 and 0.1 for 279 soils (Fig. 9c), while the highest frequency of *P*_{2000μ} is between 0.9 and 1.0 for 641 soils (Fig. 9j). In terms of the true and conventional AEVs, most of the values are in the range between 0 and 50 kPa (Fig. 9k and l), while the true AEV distribution is slightly more uniform than the conventional AEV distribution. Since the values of *P*_{2000μ} and *P*_{1000μ} do not change across different samples, they are not considered in the subsequent analyses.

Fig. 10 presents a comparison of the AEVs and true AEVs obtained for different soil classifications. As can be seen, the range of the true AEVs has shifted to larger values compared to their corresponding conventional AEVs.

The variations of the AEV and the true AEV with each input soil property are shown in graphs of Fig. 11. Also shown in the figure are the linear trend lines fitted to the data along with the coefficient of determination (COD), *R*². The COD allows for the assessment of the fitting accuracy and it is expressed as:

$$R^2 = 1 - \left[\frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y}_i)^2} \right] \tag{2}$$

$$\bar{Y}_i = \frac{1}{N} \sum_{i=1}^N Y_i \tag{3}$$

where *Y_i* and *Ŷ_i* are real and predicted output values for the *i*th output, *N* is the number of outputs, and *Ȳ_i* is the average value of the real outputs. Overall, a comparatively low fitting accuracy (*R*² < 0.177) is obtained for variations of both AEV and true AEV with input soil properties. The parameters *P*_{50μ}, *P*_{250μ}, *P*_{500μ}, *P*_{20μ}, *P*_{2μ}, *P*_{1000μ}, and bulk density, having the highest *R*² values are selected for developing the ML model and the parameters *P*_{2000μ} and porosity, having the lowest *R*² values, are omitted. This is done to avoid the complexity associated with developing the ML model.

5. Results of analysis

According to Fig. 11, the database is described by seven input parameters namely, bulk density, *P*_{2μ}, *P*_{20μ}, *P*_{50μ}, *P*_{250μ}, *P*_{500μ}, and *P*_{1000μ}, and one output parameter namely, true AEV. In order to accurately generate the predictive model, the ML method of multi-layer perceptron (MLP), which is a feedforward class of artificial neural networks (ANNs), is selected because of its simplicity and availability (comparing different ML methods is out of the scope of the current study).

ANNs are commonly used to develop practical AI frameworks to deal with complex pattern-oriented engineering problems. The nonparametric nature of ANNs permits models to be established without any prior knowledge of the distribution of the input values or presumed interactions between variables of a model as required

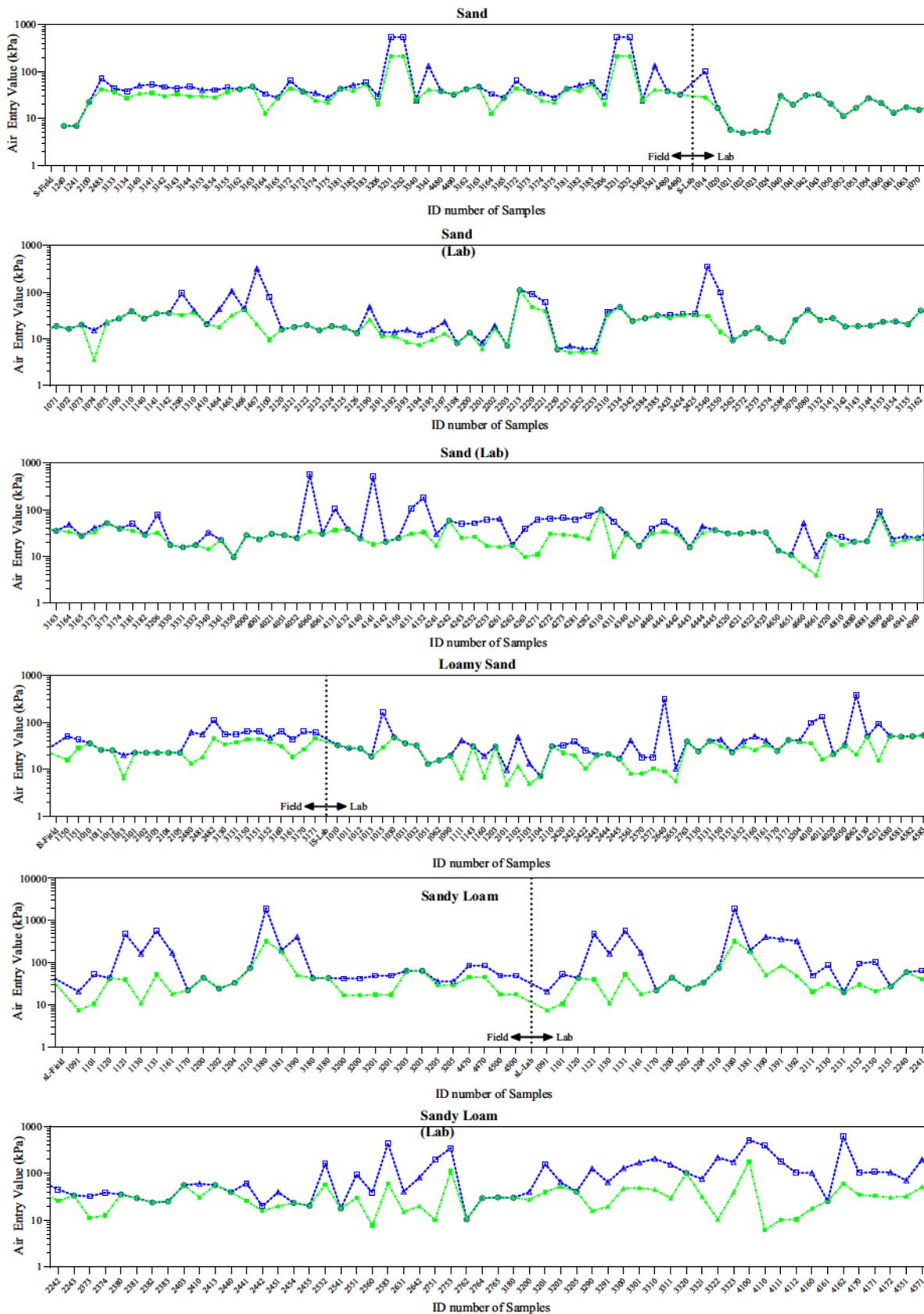


Fig. 5. AEV and true AEV for each sample of UNSODA (Blue: true AEV, Green: miscalculated AEV, Triangle: normally consolidated, Square: overconsolidated with $s_{ae} > p'_c$, and Circle: overconsolidated with $s_{ae} < p'_c$).

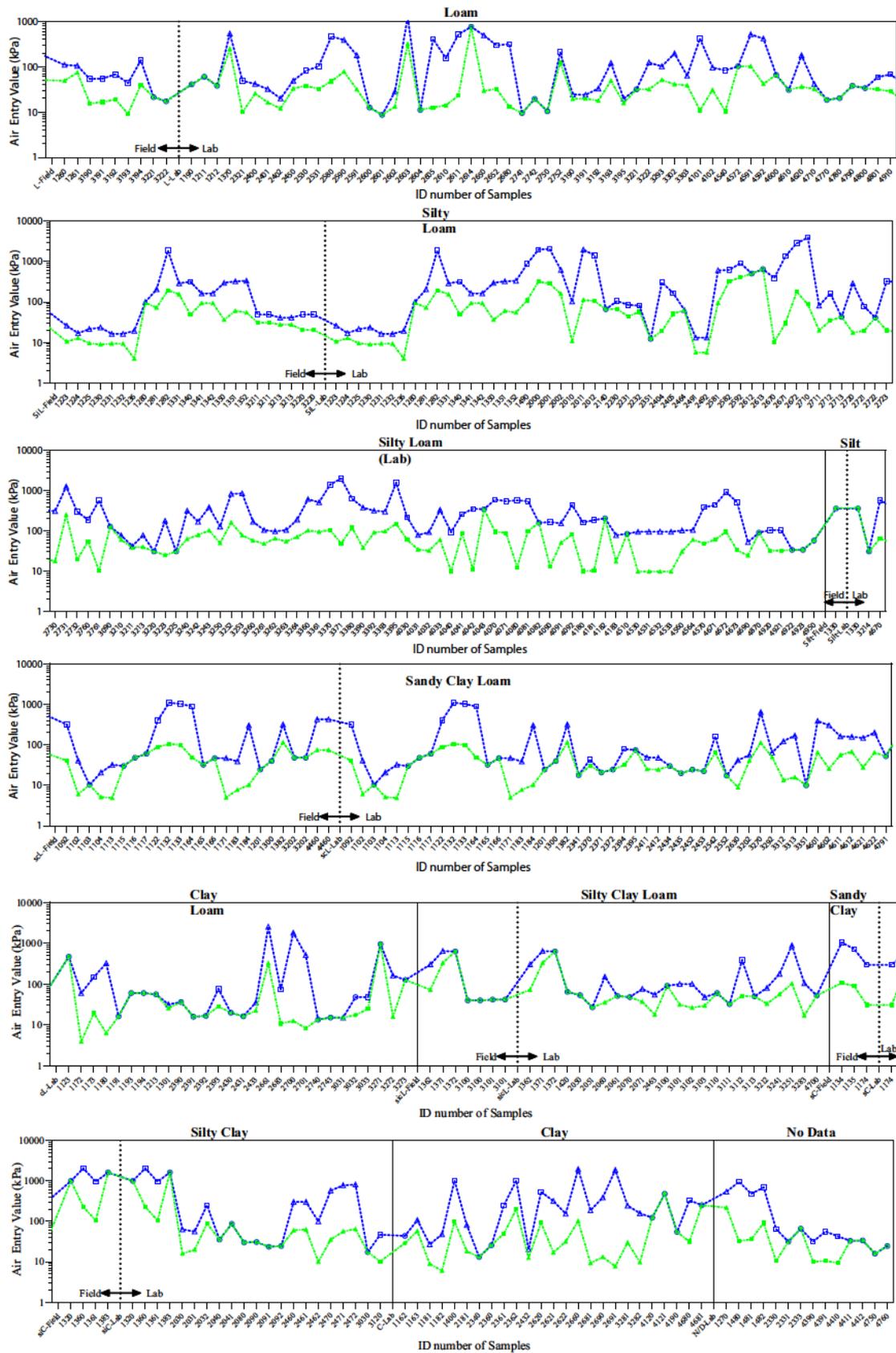


Fig. 5 (continued)

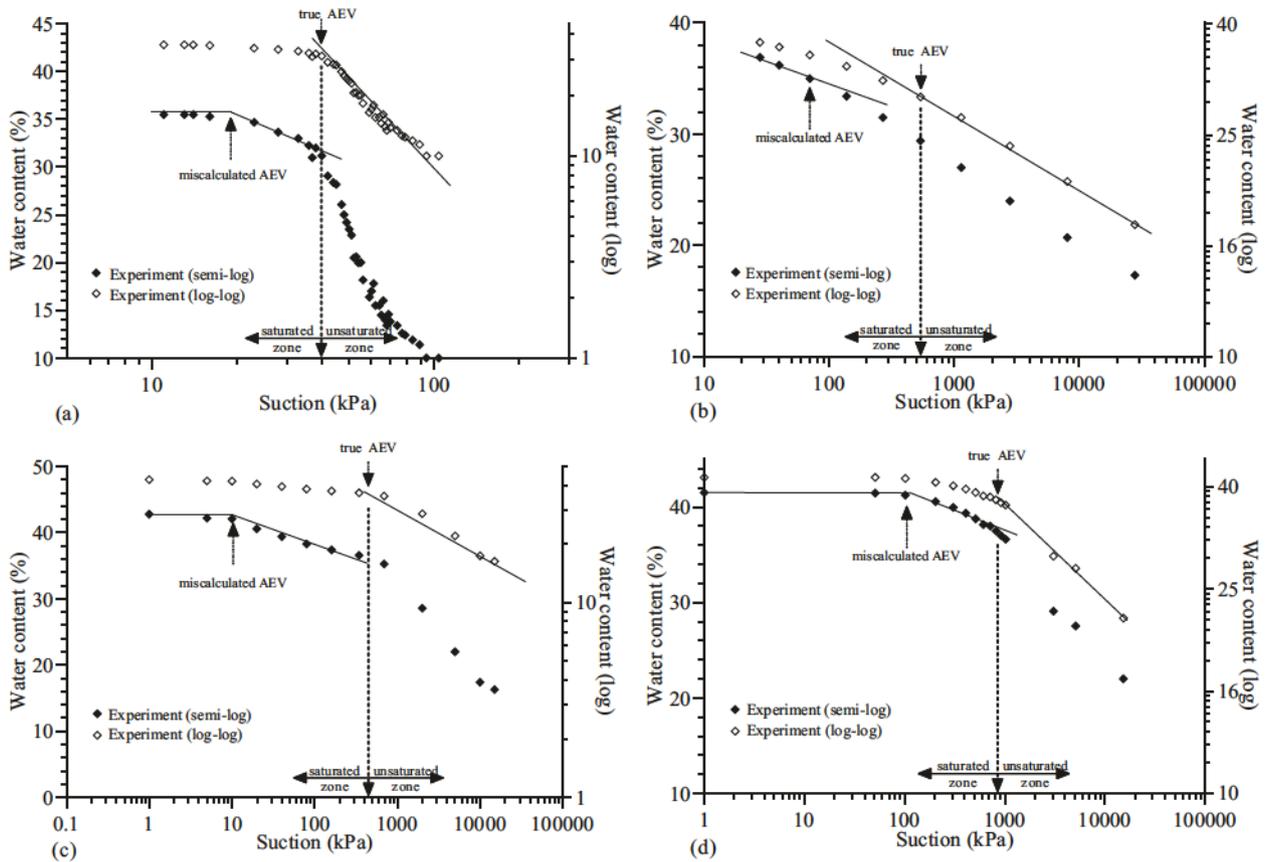


Fig. 6. Graphical determination of true and miscalculated AEVs for four soil samples in the UNSODA database (a: sample 4440, b: sample 4612, c: sample 2670, d: sample 1135).

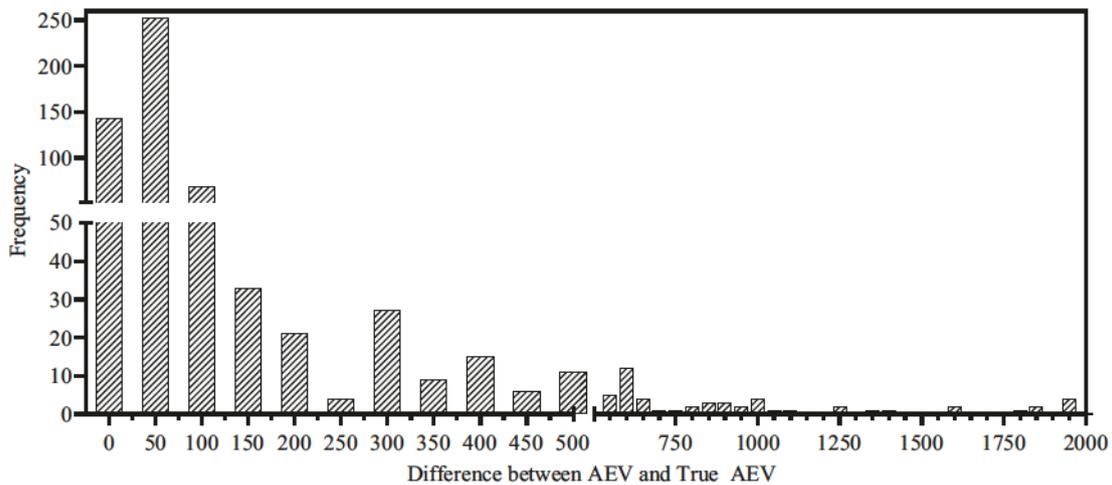


Fig. 7. Histogram of the differences between the conventional and the true AEVs.

by frequently used statistical methods (Rezaia, 2008; Walczak, 2019).

ANN can be described as a multivariate and multi-dimensional function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. It consists of an input layer of n neurons (input values), an output layer of m neurons (output values) and an arbitrary number of interior layers with a variable number of neurons, called hidden layers. The neurons are storage cells for scalar values obtained by an activation function applied to the neuron values in the previous layer (Fig. 12). To each neuron in the output

and the hidden layers, a vector of weights and a scalar bias are associated, and the value u_k^l stored by the k^{th} neuron in the l^{th} layer can be written in the form (Krenker et al. 2011):

$$u_k^l = \sigma^l \left(\sum_j \omega_{kj}^l u_j^{l-1} + b_k^l \right) \quad (4)$$

where σ^l is the activation function for the l^{th} layer, and ω_{kj}^l and b_k^l are the weights and biases respectively. Typical activation functions

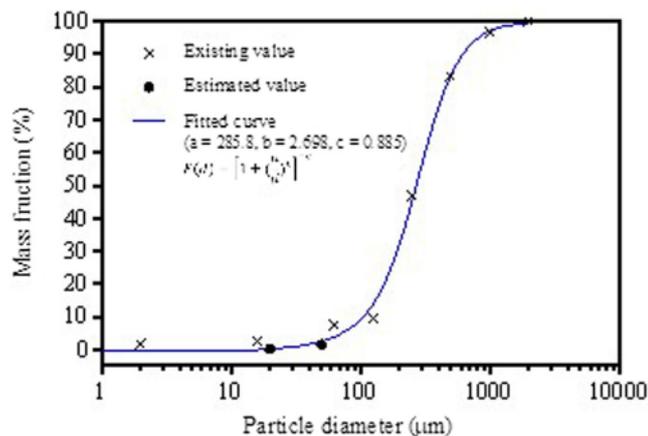


Fig. 8. Example fitted GSD curve for sample 3340.

used can be linear (i.e., $\sigma(x) = x$), or non-linear (e.g., $\sigma(x) = \tanh(x)$). Under mild assumptions on the activation functions, it can be shown that a neural network with even a single hidden layer is a universal function approximator, as given enough neurons, any continuous function on a compact domain can be approximated with arbitrary precision (Csáji, 2001). In case where more than one hidden layer is used (a so-called deep neural network), very efficient approximations can be achieved with a relatively small number of network parameters, i.e., weights and biases (Huang, 2003).

To identify the best structure of the ML model, an initial analysis is carried out to determine the optimal number of layers and neurons that can suitably represent the relationship between the inputs and output of the database. To speed up the optimization procedure, only 25% of the database is considered for this step. Fig. 13 presents a comparison of the performance of all the examined structures for one-hidden-layer and two-hidden-layer

Table 3
Descriptive statistics of all variables in the database.

Variable		Min	Max	Median	Mean	Standard deviation
Inputs	Bulk Density (g/cm ³)	0.459	1.970	1.500	1.472	0.212
	Porosity	0.264	0.915	0.444	0.461	0.093
	P _{2µ}	0.000	0.697	0.116	0.153	0.132
	P _{20µ}	0.000	0.920	0.277	0.282	0.226
	P _{50µ}	0.000	1.000	0.393	0.416	0.303
	P _{250µ}	0.000	1.000	0.853	0.783	0.225
	P _{500µ}	0.200	1.000	0.976	0.909	0.145
	P _{1000µ}	0.651	1.000	0.997	0.971	0.061
	P _{2000µ}	0.776	1.000	1.000	0.998	0.016
	Outputs	AEV (kPa)	3.548	1621.800	30.199	51.232
True AEV (kPa)		4.898	3890.400	50.119	185.427	371.372

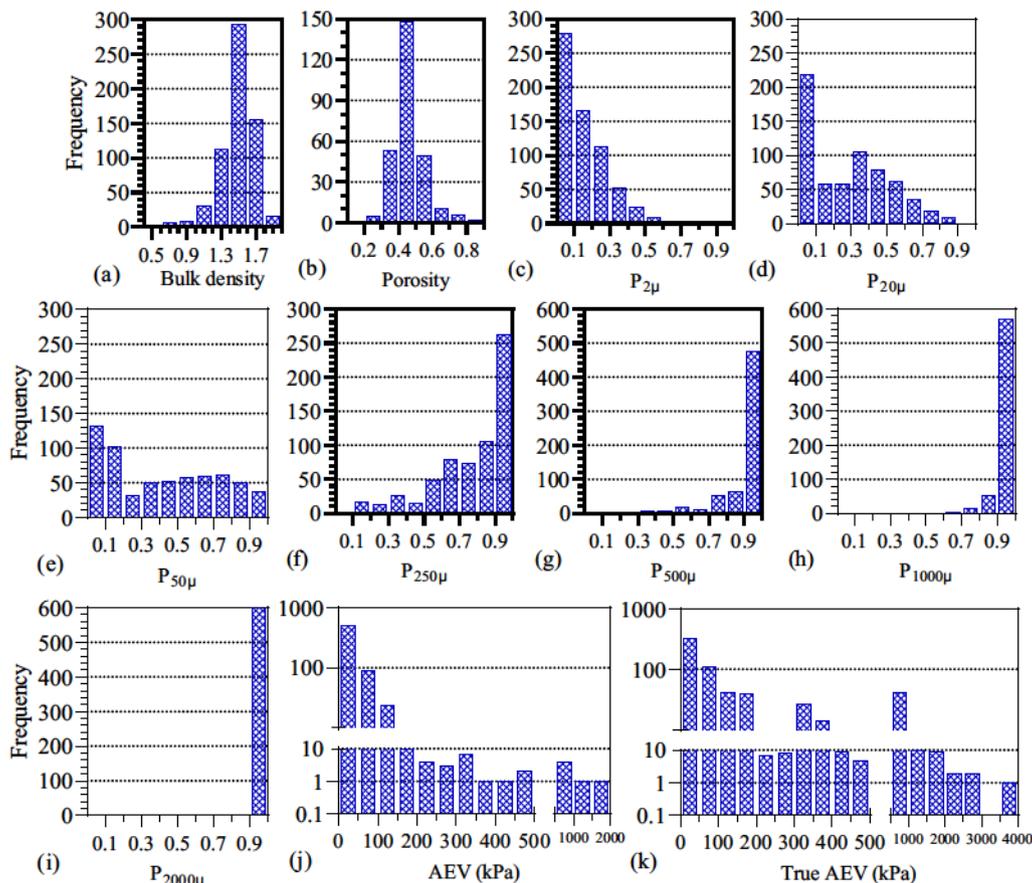


Fig. 9. Frequency histograms of the variables: (a) Bulk density, (b) Porosity, (c) P_{2µ}, (d) P_{20µ}, (e) P_{50µ}, (f) P_{250µ}, (g) P_{500µ}, (h) P_{1000µ}, (i) P_{2000µ}, (j) AEV, (k) true AEV.

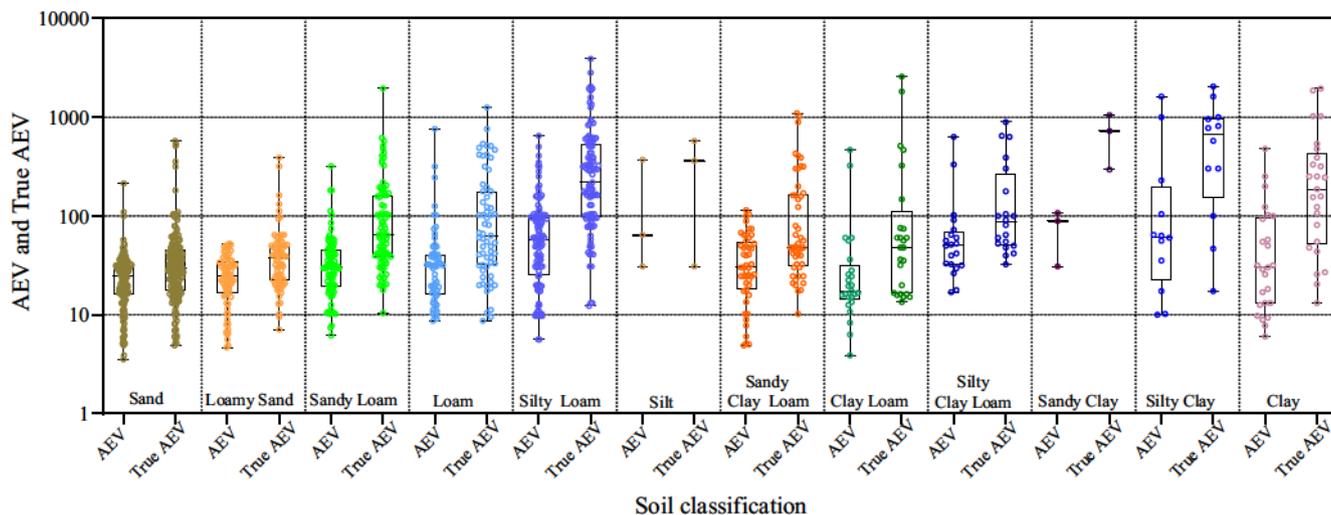


Fig. 10. Comparison of AEV and true AEV of soils with different classifications in the UNSODA database.

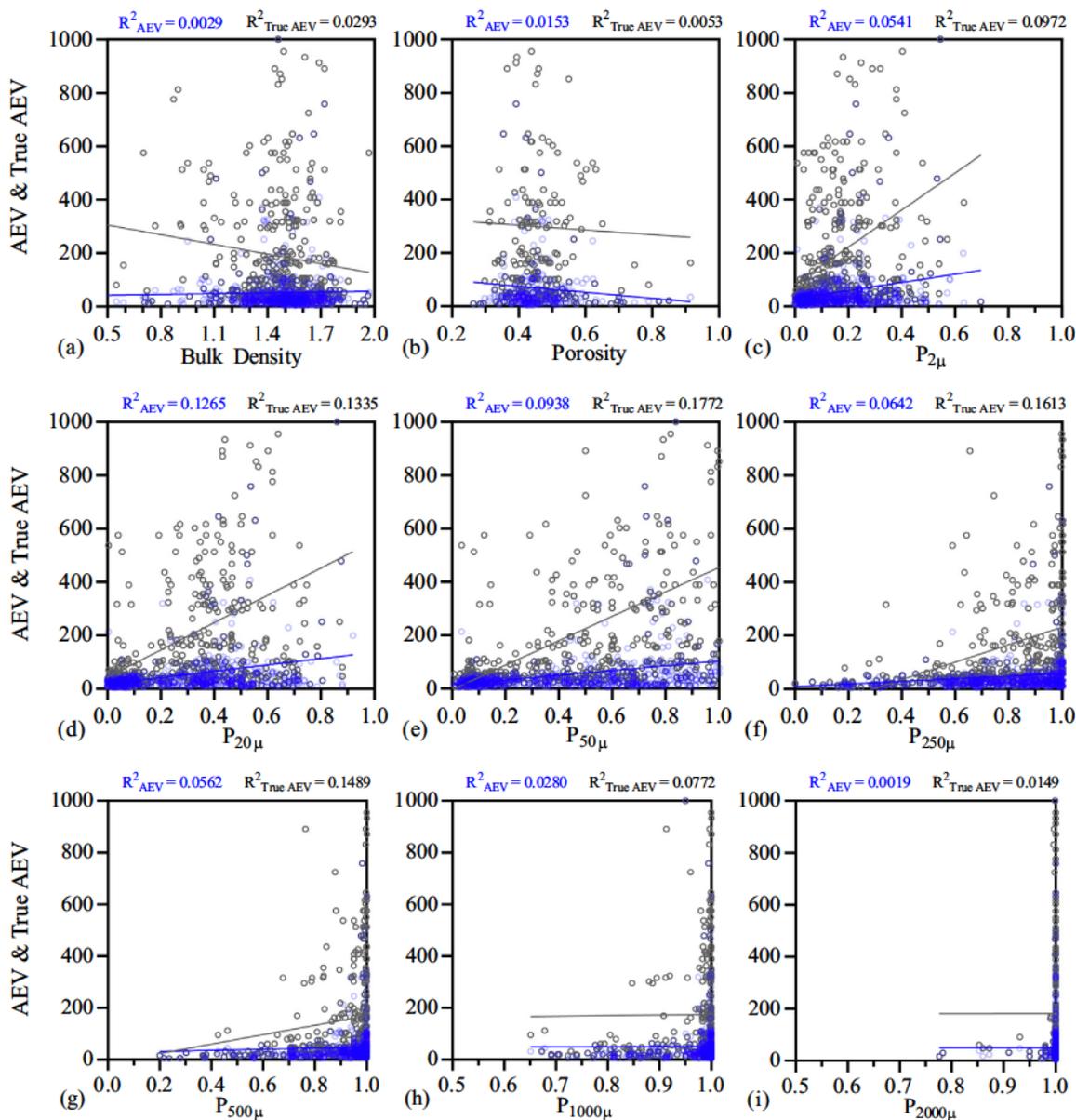


Fig. 11. Basic linear fittings between AEV and true AEV and each input soil property: (a) Bulk density, (b) Porosity, (c) $P_{2\mu}$, (d) $P_{20\mu}$, (e) $P_{50\mu}$, (f) $P_{250\mu}$, (g) $P_{500\mu}$, (h) $P_{1000\mu}$, (j) $P_{2000\mu}$.

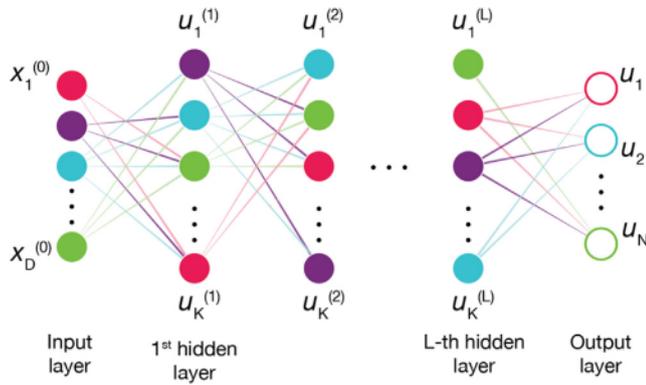


Fig. 12. A sample artificial neural network.

networks, at the end of the training process. The various architectures of ANNs are compared in terms of the RMSE given by:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (5)$$

In Fig. 13, the continuous line describes the result of the one-hidden-layer ANN, which is obviously uniquely determined by the number of neurons in the hidden layer itself. The dashed lines in this figure show the solutions for the two-hidden-layer ANNs. Each point of this graph has been calculated as the average of ten repetitions of the training process, to guarantee robustness against stochastic impacts. As shown in Fig. 13, the two-hidden-layer ANN with 5 and 23 neurons in the first and second layers provides the best performance. For cases with total number of neurons larger than the abovementioned values, there are no considerable changes in the RMSE. Therefore, the 7–5–23–1 ANN architecture is selected as the optimum ML model in this study.

In order to ensure the capabilities of the trained model to generalize its learning to unseen cases, and to prevent overfitting, only 70% of the database is used for the training process, 15% is kept for validation, and 15% is used for testing. Therefore, out of the total 644 data, 450 are used for training, 96 for validation, and 96 for testing.

Fig. 14 presents the comparison between the measured true AEVs with the ML predicted ones. It can be seen that not only the ML model exhibits a very high prediction accuracy for the training dataset, but also high-level prediction accuracies are

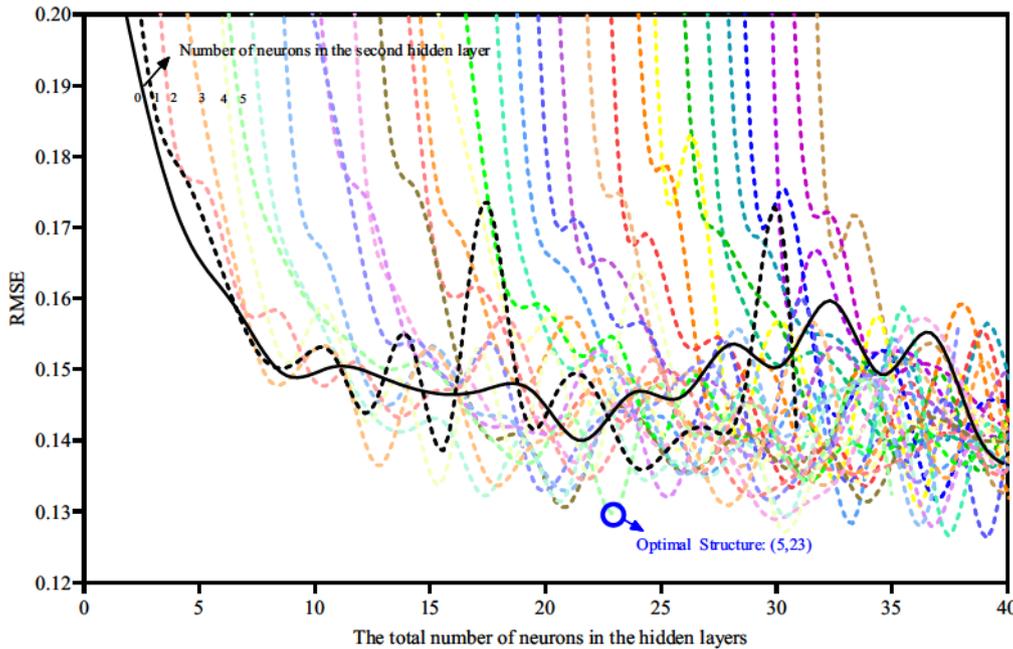


Fig. 13. Dependence of the RMSE at the end of training on the total number of neurons in the ANNs, featuring either one or two hidden layers.

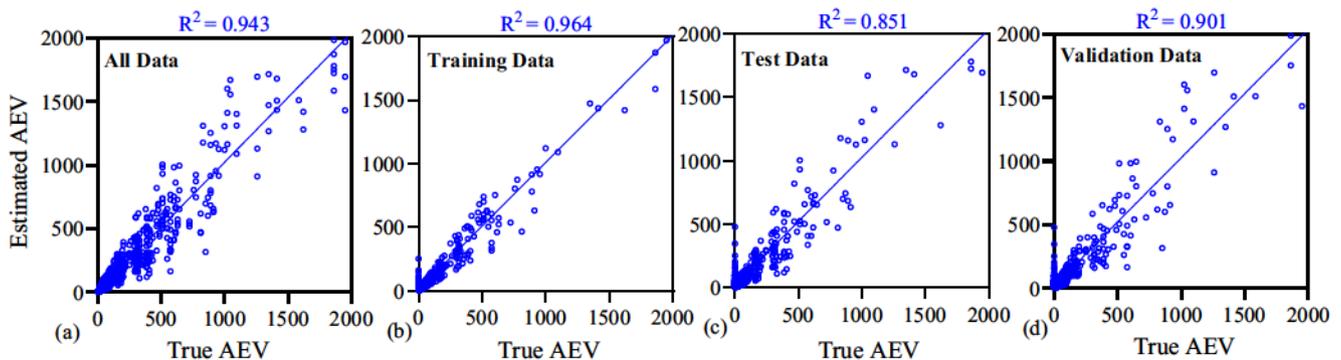


Fig. 14. Performances of the ML algorithms: parity plots showing the ML predictions against the corresponding measured values for (a) all the data; (b) training dataset; (c) testing dataset; (d) validation dataset.

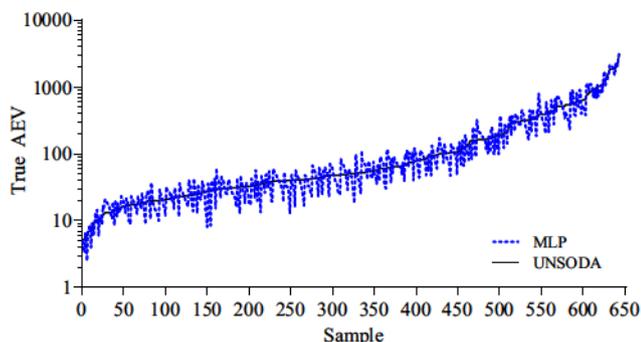


Fig. 15. Overall performance of the ML model algorithm: comparison between the measured true AEVs of all cases in the UNSODA database and the ML predicted values.

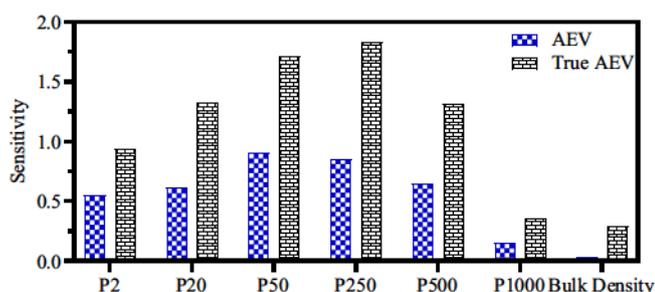


Fig. 16. Sensitivity analysis for the effects of different input parameters on the ML predicted AEVs and true AEVs.

achieved for testing and validation datasets, with the R^2 values ranging from 0.851 to 0.964 across the three datasets.

Fig. 15 presents a comparison between the measured true AEVs for all cases in the UNSODA database and the predicted true AEVs using the developed ML model. It is seen that the true AEVs estimated by the ML model match relatively well with their UNSODA counterparts; the insignificant scattering around the actual measured values are clearly in accordance with the results illustrated in Fig. 14.

6. Sensitivity analysis

To evaluate the effects of different soil parameters on soil's AEV, using the developed model a sensitivity analysis is also performed according to the method proposed by Vu-Bac et al. (2016). A sensitivity, or uncertainty, analysis quantifies the impact of all input parameters with respect to the specific output of interest, which in this study are the AEV and the true AEV. Therefore, the graphically determined AEVs from cases in the UNSODA database have been assessed for this purpose. The sensitivities of the AEV and true AEV against different independent input parameters are presented in Fig. 16 confirming that the AEV is most sensitive to $P_{50\mu}$ and $P_{250\mu}$, respectively, and the true AEV is most sensitive to $P_{250\mu}$ and $P_{50\mu}$, respectively; therefore, these 2 parameters seem to have the most effect on soil's air entry value. On the other hand, the bulk density and $P_{1000\mu}$ appear to be the less influential parameters in terms of their effect on soil's AEV.

7. Conclusions

In this paper, drawbacks in interpretation of water content-based SWRC and estimation of the AEV for a significant number of soil samples were investigated. The UNSODA database was used

and thoroughly examined for this purpose. It was shown that there are 644 SWRCs in this database that possess the required information for the targeted analysis. For these SWRCs the AEVs were obtained using the conventional method and compared with the corresponding true AEVs obtained using a consistent graphical approach proposed by Pasha et al. (2016) which considers the stress history of the samples. For the cases in the database, the differences between the conventional AEVs and their corresponding true AEVs were found to be generally in the range of 30–300 kPa, although this could reach as high as 1500 kPa. A ML model was developed to estimate the true AEVs, with the bulk density and grain size distribution factors as input parameters. The obtained results revealed that the developed ML model operates with excellent accuracy and provides practical estimations of the true AEVs with R^2 values of 0.964, 0.901 and 0.851 for the training, validation and testing datasets, respectively. Furthermore, the sensitivity analysis showed that $P_{250\mu}$ and $P_{50\mu}$ are the most influential parameters on soil's AEV. The study shows that the methodically trained ML approaches can be readily used for derivation of the true AEVs for a wide range of soils, provided appropriate information regarding the GSDs are available.

CRediT authorship contribution statement

Mohammad Sadegh Es-haghi: Conceptualization, Software, Validation, Investigation. **Mohammad Rezania:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing. **Meghdad Bagheri:** Conceptualization, Methodology, Validation, Writing – review & editing.

Declaration of Competing 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.

Acknowledgements

The second author acknowledges the financial support provided by the European Commission Research Fund for Coal and Steel (RFCS) for project MINRESCUE (Contract RFCS-RPJ-899518).

References

Alonso, E.E., Gens, A., Josa, A., 1990. A constitutive model for partially saturated soils. *Geotechnique* 40 (3), 405–430.

Alves, R.D., Gitirana Jr., G.d.F.N., Vanapalli, S.K., 2020. Advances in the modeling of the soil–water characteristic curve using pore-scale analysis. *Comput. Geotech.* 127, 103766. <https://doi.org/10.1016/j.compgeo.2020.103766>.

Amanabadi, S., Vazirinia, M., Vereecken, H., Vakiliian, K.A., Mohammadi, M.H., 2019. Comparative study of statistical, numerical and machine learning-based pedotransfer functions of water retention curve with particle size distribution data. *Eurasian Soil Sci.* 52 (12), 1555–1571.

Bagheri, M., 2018. Experimental Investigation of the Time- and Rate-Dependent Behaviour of Unsaturated Clays. University of Nottingham. Ph.D. thesis.

Bagheri, M., Rezania, M., 2022. Effect of soil moisture evaporation rate on dynamic measurement of water retention curve with high-capacity tensiometer. *Int. J. Geomech.* 22 (3), 04021301.

Bagheri, M., Mousavi Nezhad, M., Rezania, M., 2019a. A CRS oedometer cell for unsaturated and non-isothermal tests. *Geotech. Test. J.* 43 (1), 20180204. <https://doi.org/10.1520/GTJ2001-EB10.1520/GTJ20180204>.

Bagheri, M., Rezania, M., Mousavi Nezhad, M., 2018. Cavitation in high-capacity tensiometers: effect of water reservoir surface roughness. *Geotech. Res.* 5 (2), 81–95.

Bagheri, M., Rezania, M., Mousavi Nezhad, M., 2019b. Rate dependency and stress relaxation of unsaturated clays. *Int. J. Geomech.* 19 (12), 04019128.

Cheng, Z.-L., Zhou, W.-H., Garg, A., 2020a. Genetic programming model for estimating soil suction in shallow soil layers in the vicinity of a tree. *Eng. Geol.* 268, 105506. <https://doi.org/10.1016/j.enggeo.2020.105506>.

Cheng, Z.-L., Zhou, W.-H., Ding, Z., Guo, Y.-X., 2020b. Estimation of spatiotemporal response of rooted soil using a machine learning approach. *J. Zhejiang Univ.-Sci. A* 21 (6), 462–477.

- Chiu, C.F., Yan, W.M., Yuen, K.V., 2012. Estimation of water retention curve of granular soils from particle-size distribution—a Bayesian probabilistic approach. *Can. Geotech. J.* 49 (9), 1024–1035.
- Csáji, B.C., 2001. Approximation with artificial neural networks. *Facul. Sci. Etsv Lornd Univ. Hungary* 24 (48), 7.
- D'Emilio, A., Aiello, R., Consoli, S., Vanella, D., Iovino, M., 2018. Artificial neural networks for predicting the water retention curve of Sicilian agricultural soils. *Water* 10 (10), 1431.
- Fredlund, D.G., Xing, A., 1994. Equations for the soil-water characteristic curve. *Can. Geotech. J.* 31 (4), 521–532.
- Huang, G.B., 2003. Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Trans. Neural Networks* 14 (2), 274–281.
- Jain, S.K., Singh, V.P., van Genuchten, M.T., 2004. Analysis of soil water retention data using artificial neural networks. *J. Hydrol. Eng.* 9 (5), 415–420.
- Javadi, A.A., Rezaia, M., 2009a. Applications of artificial intelligence and data mining techniques in soil modeling. *Geomech. Eng.* 1 (1), 53–74.
- Javadi, A.A., Rezaia, M., 2009b. Intelligent finite element method: an evolutionary approach to constitutive modeling. *Adv. Eng. Inf.* 23 (4), 442–451.
- Krenker, A., Bešter, J., Kos, A., 2011. Introduction to the artificial neural networks. *Artificial Neural Networks: Methodological Advances and Biomedical Applications*. InTech, 1–8.
- Leij, F. J., Alves, W. J., van Genuchten, M. Th., Williams, J. R., 1996. The UNSODA unsaturated hydraulic database. USEPA, Cincinnati, Ohio. EPA/600/R-96/095.
- Li, Y., Vanapalli, S.K., 2022. Prediction of soil-water characteristic curves using two artificial intelligence (AI) models and AI aid design method for sands. *Can. Geotech. J.* 59 (1), 129–143.
- Moreira de Melo, T., Pedrollo, O.C., 2015. Artificial neural networks for estimating soil water retention curve using fitted and measured data. *Appl. Environ. Soil Sci.*
- Pagano, A., Tarantino, A., Bagheri, M., et al., 2016. An experimental investigation of the independent effect of suction and degree of saturation on very small-strain stiffness of unsaturated sand. In: *E3S Web of Conferences*, p. 14015.
- Pasha, A.Y., Khoshghalb, A., Khalili, N., 2016. Pitfalls in interpretation of gravimetric water content-based soil-water characteristic curve for deformable porous media. *Int. J. Geomech.* 16 (6), D4015004.
- Rezaia, M., 2008. Evolutionary Polynomial Regression Based Constitutive Modelling and Incorporation in Finite Element Analysis. University of Exeter. Ph.D. thesis.
- Rezaia, M., Javadi, A.A., 2007. A new genetic programming model for predicting the settlement of shallow foundations. *Can. Geotech. J.* 44 (12), 1462–1473.
- Rezaia, M., Ma, G., 2019. Stress-strain modelling of soils in drained and undrained conditions using a multi-model intelligent approach. In: *International Conference on Information Technology in Geo-Engineering*. Springer, pp. 419–428.
- Rezaia, M., Bagheri, M., Mousavi Nezhad, M., 2020. Creep and consolidation of a stiff clay under saturated and unsaturated conditions. *Can. Geotech. J.* 57 (5), 728–741.
- Russell, A.R., Khalili, N., 2006. A unified bounding surface plasticity model for unsaturated soils. *Int. J. Numer. Anal. Meth. Geomech.* 30 (3), 181–212.
- Saxton, K.E., Rawls, W.J., Romberger, J.S., Papendick, R.I., 1986. Estimating generalized soil-water characteristics from texture. *Soil Sci. Soc. Am. J.* 50 (4), 1031–1036.
- Schaap, M.G., Leij, F.J., 1998. Using neural networks to predict soil water retention and soil hydraulic conductivity. *Soil Tillage Res.* 47 (1–2), 37–42.
- Shi, C., Wang, Y.u., 2021. Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties. *Geosci. Front.* 12 (1), 339–350.
- Tarantino, A., Gallipoli, D., Augarde, C.E., De gennaro, V., Gomez, R., Laloui, L., Mancuso, C., El mountassir, G., Munoz, J.J., Pereira, J.-M., Peron, H., Pisoni, G., Romero, E., Raveendraraj, A., Rojas, J.C., Toll, D.G., Tombolato, S., Wheeler, S., 2011. Benchmark of experimental techniques for measuring and controlling suction. *Géotechnique* 61 (4), 303–312.
- Vaz, C.M.P., Ferreira, E.J., Posadas, A.D., 2020. Evaluation of models for fitting soil particle-size distribution using UNSODA and a Brazilian dataset. *Geoderma Regional* 21, e00273. <https://doi.org/10.1016/j.geodrs.2020.e00273>.
- Vu-Bac, N., Lahmer, T., Zhuang, X., Nguyen-Thoi, T., Rabczuk, T., 2016. A software framework for probabilistic sensitivity analysis for computationally expensive models. *Adv. Eng. Softw.* 100, 19–31.
- Walczak, S., 2019. Artificial neural networks. In: *Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-Computer Interaction*, IGI Global, pp. 40–53.
- Wang, H.-L., Yin, Z.-Y., Zhang, P., Jin, Y.-F., 2020a. Straightforward prediction for air-entry value of compacted soils using machine learning algorithms. *Eng. Geol.* 279, 105911. <https://doi.org/10.1016/j.enggeo.2020.105911>.
- Wang, L., Wu, C., Gu, X., Liu, H., Mei, G., Zhang, W., 2020. Probabilistic stability analysis of earth dam slope under transient seepage using multivariate adaptive regression splines. *Bull. Eng. Geol. Environ.* 79(6), 2763–2775.
- Wang, L., Wu, C., Tang, L., Zhang, W., Lacasse, S., Liu, H., Gao, L., 2020. Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method. *Acta Geotech.* 15(11), 3135–3150.
- Zhai, Q., Rahardjo, H., 2012. Determination of soil-water characteristic curve variables. *Comput. Geotech.* 42, 37–43.
- Zhai, Q., Rahardjo, H., Satyanaga, A., Dai, G., 2020. Estimation of the soil-water characteristic curve from the grain size distribution of coarse-grained soils. *Eng. Geol.* 267, 105502. <https://doi.org/10.1016/j.enggeo.2020.105502>.
- Zhang, P., Wu, H.-N., Chen, R.-P., Chan, T.H.T., 2020. Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction: a comparative study. *Tunn. Undergr. Space Technol.* 99, 103383. <https://doi.org/10.1016/j.tust.2020.103383>.
- Zhang, W., Li, H., Han, L., Chen, L., Wang, L., 2022. Slope stability prediction using ensemble learning techniques: a case study in Yunyang County, Chongqing, China. *J. Rock Mech. Geotech. Eng.* <https://doi.org/10.1016/j.jrmge.2021.12.011>.
- Zhang, W., Wu, C., Zhong, H., Li, Y., and Wang, L., 2021. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geosci. Front.* 12(1), 469–77.
- Zhou, M., Shadabfar, M., Huang, H., Leung, Y.F., Uchida, S., 2020. Meta-modelling of coupled thermo-hydro-mechanical behaviour of hydrate reservoir. *Comput. Geotech.* 128, 103848. <https://doi.org/10.1016/j.compgeo.2020.103848>.