Learning decision boundaries for cone penetration test classification

Georg H. Erharter | Simon Oberhollenzer | Anna Fankhauser | Roman Marte | Thomas Marcher

1 Institute of Rock Mechanics and Tunnelling, Graz University of Technology, Graz, Styria 8010, Austria
2 Institute of Soil Mechanics, Foundation Engineering and Computational Geotechnics, Graz University of Technology, Rechbauerstraße 12, Graz, Styria 8010, Austria

Correspondence
Georg H. Erharter, Institute of Rock Mechanics and Tunnelling, Graz University of Technology, Rechbauerstraße 12, Graz, Styria, 8010, Austria.
Email: erharter@tugraz.at

Abstract
In geotechnical field investigations, cone penetration tests (CPT) are increasingly used for ground characterization of fine-grained soils. Test results are different parameters that are typically visualized in CPT based data interpretation charts. In this paper we propose a novel methodology which is based on supervised machine learning that permits a redefinition of the boundaries within these charts to account for unique soil conditions. We train ensembles of randomly generated artificial neural networks to classify six soil types based on a database of hundreds of CPT tests from Austria and Norway. After training we combine the multiple unique solutions for this classification problem and visualize the new decision boundaries in between the soil types. The generated boundaries between soil types are comprehensible and are a step towards automatically adjusted CPT interpretation charts for specific local conditions.

1 | INTRODUCTION

Cone penetration tests (CPT) are becoming increasingly popular in geotechnical engineering and allow cost-effective and rapid in situ measurements within soils. A probe is pushed under a constant rate of 20 mm/s into the soil and records a high-resolution data profile of various parameters over the measured depth interval (see Section 2 for detailed information).

Interpretation of the resulting data is typically done by plotting it in soil type classification charts (see Figures 1 and 2; Robertson, 1991, 2009, 2016). These charts differentiate between soil types by empirically determined and continuously updated boundaries that have the goal to be as universally applicable as possible. However, it is known that albeit the charts provide a basis for CPT data interpretation, local soil conditions can deviate substantially and in practical engineering, classifications based on local expert experience often show better results than chart-based classifications. Deviations from the generally applicable charts are mainly caused by unique geological conditions, resulting from complex processes like sedimentation and consolidation (Nichols, 2009) (e.g., imagine a loosely deposited young silt in the reservoir of a power plant, compared to postglacial, silty lake deposits in the alpine foreland).

To answer for this demand of locally adjusted CPT interpretation charts, we present a novel, data driven workflow which helps to redefine and modify soil boundaries in the classical CPT charts. The goal of this paper is therefore to find well-fitting boundaries between individual soil types, solely based on the data. We see a decision boundary as “well-fitting” if it separates different soil types without being so closely fitted to the data that the boundary becomes too curvy or multiple disconnected areas of one...
soil type appear. Generating boundaries as such comes at the price of classification accuracy, as a higher accuracy is achieved the closer the decision boundary fits the data, or in other words, the more the artificial neural networks (ANNs) overfit the dataset.

The proposed workflow utilizes techniques of supervised machine learning (ML) and has new decision boundaries at its end which are generated by ensembles of ANNs (see e.g., Bishop, 2009; Géron, 2017; Raschka & Mirjalili, 2019 for in-depth information about ML). In detail, the goal is to reevaluate or confirm boundaries in the two classical CPT classification charts: normalized friction ratio \( F_r \) against the normalized tip resistance \( Q_{tip} \) after Robertson (2009) and the parameter \( U_2 \) against \( Q_{tip} \) after Schneider, Hotstream, Mayne, and Randolph (2012), which was later updated by Robertson (2016). Therefore, we train a multitude of individual ANNs with randomly defined architectures to learn to discriminate six soil types from each other. After training of all ANNs is finished, the individual results are merged, and we visualize the learned decision boundaries of all the ANNs’ features spaces of the above-mentioned CPT charts.

While the individual components of our computational framework are accepted and widely established ML methods, there is—to our knowledge—no previous application of such a data driven methodology to reevaluate soil boundaries for CPT data interpretation. In a time of ever-increasing data quantities, this study is a methodological contribution to CPT data interpretation itself that further the transition from subjective classifications towards a more objective and comprehensible soil classification.

We point out that finding supervised ML based ways to achieve high accuracies in classifying soil types is not the goal of this study and would require different approaches (see also the discussion in Section 4). Achieving highest possible classification accuracies in the 2D space of a soil classification chart would be counterproductive, as resulting soil boundaries would be less well generalizable.

The past years have seen a rapid increase in the successful application of ML for various difficult tasks such as object recognition in images (Krizhevsky, Sutskever, & Hinton, 2012), speech recognition (Hinton et al., 2012), or playing games (Silver et al., 2016). This success is largely based on deep learning—a subbranch of ML which concerns the application of deep ANNs which contain multiple processing layers with several artificial neurons each (for more information, see, e.g., Chollet, 2018; Goodfellow, Bengio, & Courville, 2016; LeCun, Bengio, & Hinton, 2015).

In comparison to other methods of geotechnical investigation, CPTs are generally well suited for ML as they are (currently) one of the few geotechnical tests that aim at high resolution data acquisition and generate high quality and high quantity data. There have been several applications of supervised and unsupervised ML for different CPT-related tasks (e.g., Goh, 1995; Kohestani, Hassanlourad, & Ardakani, 2015; Rogiers et al., 2017). See, for example, Carvalho and Ribeiro (2019) who use the two ML classification algorithms K-nearest neighbors and distance weighted nearest neighbors to replicate the CPT classifications according to Robertson (2009, 2016) and provide a comprehensive list of papers related to CPT and ML.

In Section 2 necessary background information on CPT data interpretation is provided. Section 3 presents the Machine Learning workflow—from raw data to final classification. Section 4 will show the results that were achieved and Section 5 will discuss them. A final outlook will be given in Section 6.
2 | CPT DATA INTERPRETATION

The two main parameters of CPTs are the tip resistance \( q_t \) and the sleeve friction \( f_s \), measured over depth. Additionally, the pore water pressure can be determined by means of a piezocone (CPTu) where the pore water pressure is measured above the cone at the position \( u_2 \). Furthermore, the soil’s shear wave velocity \( V_s \) can be identified for different depths using a seismic CPTu (SCPTu).

Nowadays, the application areas of cone penetration tests mainly cover the site characterization, soil profiling, and the assessment of various constitutive soil parameters by correlations. The in situ test is mainly useful in marine and lacustrine sediments, covering the grain sizes from clay to loose gravel (Schnaid, 2009).

In practical engineering, CPT-based soil behavior type (SBT) charts are mainly utilized for soil classification and soil layering. It should be noted that the SBT charts characterize the soil according to their “behavior.” This “behavior” results from the soil’s grain size distribution, but is also highly influenced by processes of deposition, consolidation and its stress history.

Today the normalized SBT chart according to Robertson (2009) and Schneider et al. (2012) are mostly used in practical engineering. Robertson (2009) defines the soil by means of normalized tip resistance \( Q_{tn} \) and the normalized friction ratio \( F_r \) (see Figure 1):

\[
Q_{tn} = \left[ \frac{(q_t - \sigma_{vo})}{p_a} \right] \left( \frac{p_a}{\sigma'_{vo}} \right)^n \tag{1}
\]

\[
F_r = \left[ \frac{f_s}{(q_t - \sigma_{vo})} \right] \times 100\% \tag{2}
\]

where \( q_t = q_c + u_2 \cdot (1 - a) \) represents the tip resistance corrected for water effects; \( u_2 \) is the pore water pressure measured above the cone; \( a \) is the cone area ratio determined by means of laboratory tests or calibration measures; \( p_a \) is the atmospheric reference pressure; \( \sigma_{vo} \) and \( \sigma'_{vo} \) represent the total and effective vertical in situ stress respectively; \( n \) is a variable stress component.

As shown in Figure 1, Robertson (2009) classifies the soil into the following nine groups: sensitive fine-grained, organic, clay, silt-mixtures, sand-mixtures, sand, gravelly sand to sand, very stiff sand to clayey sand, and very stiff fine-grained. Additionally, a sector for normally consolidated soils is defined by two blue dashed lines.

Schneider et al. (2012) suggested a soil behavior type chart based on the normalized tip resistance \( Q_{tn} \) and \( U_2 \) (where \( U_2 = (u_2 - u_0)/\sigma'_{vo} \)). As shown in Figure 2, Robertson (2016) proposed an updated Schneider et al. (2012) chart based on \( Q_{tn} \) and \( U_2 \). The updated soil behavior type chart defines the following soil groups: CCS (clay-like-contractive-sensitive), CC (clay-like-contractive), TC (transitional-contractive), and SD (sand-like dilative).

Besides using CPT tests directly for soil classification, it is common practice to use them as an addition to the more expensive core drillings. In this case core drillings with core logging and grain size analyses are performed to characterize the soil stratigraphy.

3 | MACHINE LEARNING PIPELINE

In this paper the term “feature” denotes a parameter measured during a CPT test or computed based on the in situ measurements (e.g., \( q_t, f_s, Q_{tn}, \) or \( F_r \) are features).

The chart \( Q_{tn} \) versus \( F_r \) based on Robertson (2009) will mostly be used for explanations, however the given processes are exactly the same for the chart \( Q_{tn} \) versus \( U_2 \). Figure 3 shows a graphical representation of the individual steps of the proposed methodology with: (1) raw data (Section 3.1), (2) data preprocessing (Section 3.2), (3) ANN design and training (Sections 3.3 and 3.4), (4) ensemble classification and decision boundary visualization (Section 3.5), and finally, visualization of areas with data (Section 3.6).

3.1 | Raw data

In situ tests (CPT, CPTu, and SCPTu), executed within Austria and Norway, represent the data basis of the current article and were executed by the Premstaller Geotechnik ZT GmbH as well as the Norwegian Geotechnical Institute, respectively. In a first step, 1,490 in situ tests were implemented to a QGIS database (Austria: 757 CPT, 612 CPTu, and 97 SCPTu; Norway: 24 CPTu). For this study only CPTu and SCPTu tests were used, due to the influence of \( u_2 \) on \( Q_{tn}, F_r, \) and \( U_2 \). An overview of the executed test sites is presented in Figure 4.

The in situ tests were executed within the Austrian basins of Salzburg and Zell as well as the region of Flachgau. Therefore, in addition to a holistic interpretation of all the in situ measurements, local differences were elaborated using the three mentioned areas (Salzburg basin, Zell basin, and region of Flachgau).

The four Norwegian test sites are located in the southern part of Norway and complement the Austrian dataset. The Norwegian dataset comprises CPTu from selected test sites featuring ground conditions consisting predominately of sand, silt, clay, or quick clay.

In order to enable an interpretation of the in situ tests based on the grain size distribution, core drillings (in combination with a soil description) executed within a maximum distance of approximately 50 m to the in situ tests.
were included to the overall database (Austria: 160 core drillings; Norway: eight core drillings). The soil classification from the drillings was assigned to the single in situ tests and lastly, the soil descriptions were subdivided into six groups based on EN ISO 14688-1 (Österreichisches Normungsinstitut, 2019) as shown in Table 1.

For a detailed description of the Austrian database reference should be made to Oberhollenzer, Fankhauser, Marte, Tschuchnigg, and Premstaller (2020). The Norwegian CPTs were recorded as part of the NGTS project (The Research Council of Norway, Norwegian GeoTest Sites (NGTS): Infrastructure Grant No. 245650/F50) and information on the four used test sites can be found in the following papers: Tiller Flotten (L’Heureux, Lindgård, & Emdal, 2019), Halden (Blaker, Carroll, Paniagua, DeGroot, & L’Heureux, 2019), Onsøy (Gundersen, Hansen, Lunne, L’Heureux, & Strandvik, 2019), and Øysand (Quinteros et al., 2019).

### 3.2 Preprocessing

Preprocessing is necessary to bring the data into a fitting format for ANN training and classification. Table 2 gives an overview of the steps of preprocessing in the order that they

### Table 1 Classification of the soil according to the grain size distribution (Oberhollenzer et al., 2020)

| Group | Grain-size distribution |
|-------|-------------------------|
| 1     | Coarse sand to gravel (CSa → Gr) |
| 2     | Peat (Or) |
| 3     | Fine to coarse sand (FSa → CSa) |
| 4     | Fine-sandy silt to silty fine sand (fsaSi → siFSa) |
| 5     | Clayey silt to fine-sandy silt (cl-Si → fsaSi) |
| 6     | Silt-clay mixtures to clayey silt (Si/Cl → clSi) |

### Table 2 Steps of data preprocessing. Steps marked with a * will be explained in more detail in the text

| Step | Description |
|------|-------------|
| 1    | Manual assignment of core logs (i.e., grain size description) to the CPT data according to the respective depth; = label assignment |
| 2    | Compute the features “log Q_{in}” and “log F_r” from the original Q_{in} and F_r |
| 3    | Delete data points that are below/above the following ranges, as they are considered to be outliers: [−0.3 < log Q_{in} > 7], [−2.5(\log F_r)2.8], [−2.8(U_2)11.5] |
| 4    | Scale data between 0 and 1 |
| 5*   | Balance the soil type classes of the dataset by applying the SMOTE algorithm |
| 6*   | Train–test split: randomly sample 10% of the data for testing purposes |
| 7*   | One-hot encode labels of the dataset |
were applied. The order of these steps may be changed to some extent. However, some order must be kept to achieve reasonable results:

- step 3 must be done after step 2, as otherwise there are no respective features,
- step 4 must be done after step 3, as outliers might lead to erroneous data scaling and information loss,
- step 5 must be done after step 3 (outliers might cause problems), but could also be done before step 4 as 0–1 scaling of the balanced dataset would result in the same ranges of data,
- step 6 could theoretically be done after step 3, but as the training and the test data set have to be scaled and balanced in the exact same way it would unnecessarily complicate the whole workflow to do so,
- step 7 must be done after step 1, as otherwise there are no labels.

### 3.2.1 Synthetic Minority Over-sampling Technique

The original dataset shows an imbalanced distribution of the six soil types (STs) where ST 2 is highly underrepresented (~5% of the whole dataset) and the STs 4 and 6 are overrepresented (see Figure 5).

As neither underground conditions, nor projects where CPTs are performed are homogeneously distributed in “real world” applications, it is necessary to find a way of balancing the dataset. Otherwise, a ML algorithm might learn that a high classification accuracy is achievable by always classifying the overrepresented classes. With six soil types, each class would make up 16.6 % (i.e., 100/6) in an ideally balanced dataset.

To balance the dataset, the SMOTE algorithm (Synthetic Minority Over-sampling Technique after Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was used. SMOTE is an oversampling technique that synthesizes unique samples of the underrepresented class by interpolating along lines between the “real” datapoints (Figure 6). Oversampling means that the number of instances of the underrepresented class is increased, in contrast to undersampling where the number of instances of the overrepresented class is decreased.

Figure 7 shows the dataset before and after SMOTE was applied and Figure 5 shows each class’ percentage share of the whole dataset before (light gray bars) and after (dark gray bars) SMOTE was applied to counter the imbalanced distribution of classes. Note how in Figure 7 the underrepresented classes become more pronounced, the overall amount of data increases (from 325,063 to 490,914 datapoints), but also that the median values of each class remain at the same positions.

### 3.2.2 Train-Test Splitting

After balancing the classes, the dataset was split into a training and testing dataset for every individually trained ANN. Ninety percent of the datapoints are used for training and 10% of them are randomly sampled for testing purposes. The train-test split is necessary as the test dataset is used to validate the training success and to guarantee that the ANN is not memorizing the training data. As given above, the final classification/decision boundaries are the product of multiple ANNs which are independently trained. We used a new train-test split of the dataset for each individual training run in order to avoid that the ANNs overfit one particular subset of data. The use of different training and testing datasets for each training run also serves the purpose of cross validation (James, Witten, Hastie, & Tibshirani, 2017).

### 3.2.3 One-hot encoding of labels

One-hot encoding is a technique in ML that is used to deal with nominal–categorical data (Raschka & Mirjalili, 2019). If a dataset contains $N$ different class labels, they are each converted into a binary vector of length $N$ where every class has its own fixed position. The target class label is represented as 1 and the other classes as 0 (Harris & Harris, 2016, p. 129), (e.g., class $3 = [0, 0, 1, 0, 0, 0]$).
3.3 | Models

ANNs are used as classifiers, which are known to be inherently unstable. This means that if two ANNs with identical architectures are independently trained to fulfill a certain task, two different solutions will be found (e.g., Cunningham, Carney, and Jacob, 2000). Instability may arise for different reasons like random weight initialization, insufficient training data, ANNs getting “stuck” in local minima during optimization, and so on (see Dietterich, 2000).

We use this instability to find robust decision boundaries by combining a multitude of individual classifications. To further increase the variability of the individual results, we generated individual ANN architectures randomly as suggested by Chollet (2018, p. 266). The basic idea is that each classification is seen as one “expert opinion” and thus the final result represents an average of many different “opinions.” The final boundaries between soil types are created by the unweighted average of 250 different classifications. The more classifications we combined, the less noisy the boundaries became, and we found 250 to be an amount of classifications that produces mostly noise-free boundaries (see Figure 8). Combining more than 250 classifications did not lead to improvements. Multilayer perceptrons (MLP; i.e., standard feedforward ANNs) were used as classifiers and to further increase the diversity of the individual classifications, each MLP’s architecture was randomly designed (within boundaries, see below).

All deployed MLPs have an input layer with two neurons for the two input features (i.e., either $Q_{tn}$ and $F_r$ or...
The output layer consists of six neurons where each neuron corresponds to one of the six one-hot encoded class labels (see previous Section). The output layer uses a softmax function (Equation 3) which transforms a vector \( a \) of length \( K \) (i.e., 6) in a way that all elements are scaled between 0 and 1 and their sum adds up to 1 (i.e., final output vector \( \hat{y} \)) (Bishop, 2009, p. 198).

\[
\hat{y}(a) = \frac{e^{a_j}}{\sum_{k=1}^{K} e^{a_k}} \tag{3}
\]

The number and size of hidden layers was however randomly chosen, where each MLP has one to three hidden layers and each hidden layer contains 2–10 neurons. The smallest possible MLP therefore contains only one hidden layer with two neurons and the biggest possible MLP three hidden layers with 10 neurons each.

The neurons in the hidden layers deploy rectified linear unit (ReLU) activation functions (Hahnloser, Sarpeshkar, Mahowald, Douglas, & Seung, 2000). ReLU (Equation 4) has shown to be more efficient in training deep ANNs in comparison to the formerly popular sigmoid activation functions (Glorot, Bordes, & Bengio, 2011) and is currently widely used in different ANN architectures.

\[
f(x) = \max(0, x) \tag{4}
\]

The Python library Keras (Chollet and others, 2015) with a Tensorflow backend (Abadi et al., 2015) is used to design and train the MLPs. The “Adam” optimization algorithm (Kingma & Ba, 2014) was used for all MLPs as it has also shown successful applications in many state of the art projects and “categorical cross-entropy” (Murphy, 2012) was used as a loss function—both with Keras’ default configuration.

3.4 | Neural network training

Despite using a small overall ANN size (see previous Section) we use “early stopping” during training to mitigate overfitting. Training is aborted once the loss (i.e., categorical cross-entropy) does not decrease for six consecutive epochs and the last best scoring ANN is saved and used for the classification. Additionally, to the loss, the training and test data’s classification accuracy is logged during training.

The whole process of training 250 independent ANNs to find robust decision boundaries, was done four times for both charts (\( Q_{tn} \) and \( F_r \) or \( Q_{tn} \) and \( U_2 \)) resulting in eight total runs: two times with the whole dataset (i.e., all CPT tests), and two times for each subset of data containing only samples from the main sedimentary basins of Salzburg and Zell as well as the region of Flachgau (the dataset’s three biggest basins).

As given in Section 3.2.2, each individually trained ANN used a different split of a training–testing dataset, but most classifications reached an overall accuracy of around 50% (Figure 9 bottom row). The accuracy of around 50% is consistent throughout the eight different subsets of the data. As given above, a higher accuracy could be achieved by bigger networks and longer training, but this would undermine the goal of finding robust and generalizable decision boundaries. Furthermore, it can be seen in Figure 7, that the individual data points of the soil types are highly overlapping which reduces the achievable accuracy as the goal is to do the classification in the 2D space of the classical Robertson charts. Therefore no “clean” and highly accurate separation between soil types can be found in the 2D space (see also the discussion on accuracy in Section 5).

We furthermore observed that during all training runs, several of the ANNs became “stuck” at an accuracy of around 16.6% (see Figure 9). Inspecting the individual
FIGURE 9 Exemplary test–data loss (i.e., categorical cross-entropy) and accuracy that were recorded during independently training 250 Multilayer Perceptrons to classify the six soil types based on log $Q_{tn}$ and log $F_r$ from all CPT tests. Note how several ANNs become “stuck” at an accuracy of $\sim 16.6\%$, which means that they only classify one single soil type

results, we found that these ANNs only classified one single soil type which consequently led to an overall accuracy of around $16.6\%$ due to the perfectly balanced dataset (see Section 3.2.1). Due to the above described “early stopping,” training of these ANNs was automatically aborted at an early stage due to a stagnating loss. However, in comparison to all classifications, only around 10% of the classifications show this problem and the “one soil type” classifications are randomly spread across all six soil types. Consequently, this is not introducing categorical bias in the final result and we did not take any countermeasures.

3.5 | Ensemble classification and certainty estimation

To visualize decision boundaries of a trained ANN, one can let the ANN classify a range of datapoints in a certain interval and as a result it becomes visible which “decision regions” (Raschka & Mirjalili, 2019) the ANN delineates. To visualize the decision boundaries of each ANN of this study, we let the trained ANNs classify all datapoints of the two 2D arrays within the intervals $\{0.1 \leq F_r \leq 10\}$ and $\{1 \leq Q_{tn} \leq 1000\}$, respectively $\{-2 \leq U2 \leq 20\}$ and $\{1 \leq Q_{tn} \leq 1000\}$ (in accordance with the intervals of the Robertson (2009, 2016) diagrams). To generate a high-quality visualization of the decision boundaries, we chose a resolution of $1,024 \times 1,024$ datapoints. The final result of each ANN’s classification is therefore a hypermatrix of the shape $1,024 \times 1,024 \times 6$ with six channels—one channel for each of the six soil types (this is comparable to an RGB image which is a hypermatrix with three channels containing the color information red, green, and blue).

With the goal to get the “average expert opinion” of where the decision boundaries are supposed to be, we combined the 250 classifications of the independently trained ANNs, by taking the arithmetic mean of all classifications. We call this an “ensemble classification” as it is inspired by ensemble machine learning (see e.g., Breiman, 1996; Cherkauer, 1996; Dietterich, 2000). In Figure 10, the whole process of ensemble classification and certainty estimation is visualized.

To illustrate this process, let a single point be classified independently five times:

ANN1: $[0, 0, 1, 0, 0, 0] \rightarrow \text{class} 3$
ANN2: $[0, 0, 0, 1, 0, 0] \rightarrow \text{class} 4$
ANN3: $[0, 0, 1, 0, 0, 0] \rightarrow \text{class} 3$
ANN4: $[0, 0, 0, 0, 1, 0] \rightarrow \text{class} 5$
ANN5: $[0, 0, 1, 0, 0, 0] \rightarrow \text{class} 3$

Then the average classification would be class 3 with the values $[0, 0, 0.6, 0.2, 0.2, 0]$. Following this procedure, the maximum value of the result (0.6 in this example) indicates not only the most probable class but is also an indicator for how “certain” the different ANNs are about their decision. The maximum, achievable certainty is therefore 1, if all ANNs agree, respectively the minimum possible certainty is $\frac{1}{n_{\text{classes}}}$, where $n_{\text{classes}}$ denotes the total number of classes (i.e., 6 classes in this case and a minimum certainty of 0.16). To make the certainty estimation comparable to other cases with different numbers of classes we developed Equation (5) that computes the certainty ($c$) for a single datapoint as the maximum value of the average classification, scaled between 0 and 1.

$$c = \max \left( \frac{\sum_{i=1}^{n} y_i}{n} \right) - \frac{1}{n_{\text{classes}}}$$

The certainty computed after Equation (5) was then used to create an “ANN certainty map” that visualizes which regions of the plot have the highest uncertainty respectively in which regions of the plot the ANNs show the highest disagreements.

3.6 | Visualization of areas with data

Each ANN’s individual classification, as well as the ensemble classifications and the uncertainty maps cover the whole ranges of values in the Robertson charts. Albeit the charts are completely covered, only areas that contain datapoints should be considered for further interpretation. As the datapoints of the given dataset do however not cover
FIGURE 10 3D visualization of how an ensemble classification and a certainty map is created from individual ANN classifications

FIGURE 11 Different selected examples of individual classifications of the six soil types with the features $F_r$ and $Q_{tn}$ on the whole CPT dataset

the whole region of the Robertson chart (see e.g., Figure 7), a gaussian kernel density estimation was done to visualize the areas where no datapoints are present. Outside of these areas, the classifications are not considered.

4 | RESULTS

Inspecting individual classifications, we observe that the strategy to increase the variability of the results worked well. The ANNs generated a multitude of unique results, ranging from very simple to very complex solutions for the given classification problem. For example, we refer to a simple solution as a linear separation of individual soil types and to a complex solution as a nonlinear separation of soil types, eventually even including multiple, disconnected areas per class.

Figure 11 shows three selected examples with different complexities of individual ANN classifications of the whole dataset in the 2D feature space $Q_{tn}$ versus $F_r$: The left—very simple—classification was created by an ANN with two hidden layers with four and two neurons each and reached a classification accuracy of 28.99%. The middle classification was created by an ANN with two hidden layers with five and eight neurons each and reached an accuracy of 44.61%. The right classification was made by an ANN with three hidden layers with nine, ten, and seven neurons each and reached an accuracy of 48.04%.

Independently which subset (i.e., basin) of the data was used, we generally observed that the bigger ANNs (in terms
of numbers of layers and neurons) find more complex classifications and achieve higher accuracies. However, as given above and as it can be seen in Figure 11 (right), a higher accuracy does not necessarily lead to better suited soil boundaries but indicates a better “fitting” of the ANN to the given dataset. Except for the classifications, where an ANN was “stuck” on one class (see Section 2.4), each individual classification is a unique. Whether an ANN becomes “stuck” on one class seems to be unrelated to the size of the ANNs since both small and big ANNs became “stuck” on classes.

In Figure 12 and Figure 13 the ensemble classifications of the $Q_{tn}$ versus $F_r$ chart and the $Q_{tn}$ versus $U_2$ chart and their certainty maps are shown. In both figures, from top to bottom, the rows represent the results for the holistic dataset, the subset of the data for the Salzburg basin, Zell basin, and the region of Flachgau.

It is visible that the areas where there is low certainty (i.e., areas where the ANNs are not in good agreement) are located outside of areas with sufficient data (see Section 3.6) and at boundaries between classes. The new soil type boundaries are in good accordance with the ideas behind the Robertson charts (e.g., decreasing grain sizes from one side of the chart to another). However, in detail the new soil boundaries are different from the Robertson charts which shows that the goal of creating locally adjusted charts was reached.

The $Q_{tn}$ versus $F_r$ chart according to Robertson (2009) presents a decreasing trend with respect to the grain size distribution from the top left corner of the chart (area 7 = gravelly sand to sand) toward the bottom right (area 3 = clay). Similar patterns are reached using the ANN classifications (Figure 12). In all cases, ST 1 and in one case also ST 3 are located in the upper left corner, which both represent sand dominated classes. Going from top left to bottom right, the most comprehensible succession from coarse to fine was reached within the Salzburg basin. The trend is in good agreement with the soil behavior type chart according to Robertson (2009), whereby the location of the boundaries differ. Except for the Zell basin (third row), ST 2 (i.e., peat/organic sediments) is never located at the very lower right corner but rather located in areas of medium $Q_{tn}$ (10–100) and high $F_r$ (> 1). While the results within the Salzburg basin show a well comprehensible succession, the region of Flachgau and the Zell basin are less comprehensible and the areas for ST 4, 5, and 6 (i.e., the finer grained STs) are partly located in ambiguous locations. The in situ measurements of the Zell basin lead to the least comprehensible results using the $Q_{tn}$ versus $F_r$ chart as the generated classification is patchy and some STs appear in multiple locations.

As the Robertson (2009) charts discriminate nine soil types and the used dataset consists of six classes, a direct evaluation of one system against the other is not possible. Qualitatively it can however be observed that the learned decision boundaries fit better to the given dataset: whereas according to Robertson (2009) organic sediments should be situated in the outermost lower right corner (see Figure 1), in this dataset, organic sediments can be found in the area of high $F_r$ and in the medium range of $Q_{tn}$ (see e.g., the median of ST 2 in Figure 7). In the Robertson (2009) chart these sediments would be classified as clay, silt-mixtures or very stiff fine-grained soils (numbers 3, 4, and, 9 in Figure 1, respectively). The position of the decision boundary of the organic sediments in the top row of Figure 12 (i.e., orange colored) exemplifies how the learned boundaries are an improvement to the existing ones as they fit much better to this dataset.

Concerning the decision boundaries for the soil behavior type chart after Schneider et al. (2012) and updated by Robertson (2016) (i.e., $Q_{tn}$ versus $U_2$): According to Robertson (2016), sand like—dilative soils (SD) are supposed to show up in the upper left corner of the diagram (i.e., low $U_2$ and high $Q_{tn}$). The transitional zone (TC) and fine-grained sediments (i.e., clay like—CC and clay like sensitive CCS) are characterized by smaller $Q_{tn}$ as well as higher $U_2$ values (see Figure 2). As shown in Figure 13, the characteristics of the learned decision boundaries are in good agreement with the trend according to Robertson (2016). For example, going from high to low $Q_{tn}$ values (at low $U_2$ values) in the decision boundary chart of the data subset of Salzburg (second row in Figure 13), the succession ST 1 $\rightarrow$ ST 3 $\rightarrow$ ST 4 $\rightarrow$ ST 2 $\rightarrow$ ST 5 $\rightarrow$ ST 6 was reached and corresponds to a transition from coarse to fine. In the area where sufficient data are present, a similar succession is observed in the learned decision boundaries for the Zell basin. The chart for the region of Flachgau also shows comprehensible results, as well as little disagreement between the ANNs in the area where there is data. The STs 1, 2, and 3 are located at the top of the chart as well as elongated from top to bottom with an increasing $U_2$ from coarse to finer soil types. The bottom right part of the chart consists of the fine-grained soil type 6 which is also in good accordance to the Robertson charts.

5 | DISCUSSION

The proposed methodology is a new approach to find classification boundaries between different soil types based on $Q_{tn}, F_r$, and $U_2$. The new boundaries are “locally adjusted” as they represent the soil types’ distribution on the Robertson charts for a defined geographical area. The presented concept therefore is an improvement of the soil behavior type charts which are widely used in practical engineering and do not lead to sufficient results for
Final results of the learned decision boundaries in the $Q_{tn}$ versus $F_r$ chart (left column) and ANN certainty maps (right column). The soil type boundaries of Robertson (2009) and class numbers are plotted in the uppermost row as reference. Black dashed lines delimit areas of sufficient data point density. From top to bottom the rows present the result for: holistic dataset, Salzburg basin, Zell basin, and the region of Flachgau.
FIGURE 13  Final results of the learned decision boundaries in the $Q_n$ versus $U_2$ chart. Soil type boundaries and class symbols after Robertson (2016) are also plotted in the uppermost row as reference. Other symbology and row arrangement is identical to Figure 12.
transitional (i.e., silt dominated) soils (see Oberhollenzer et al., 2020).

Quantifying the proposed framework’s accuracy is however not directly possible, as the classification accuracy is a measure that can be used to monitor the state of the training progress but maximizing it as the overall target leads to worse soil boundaries and overfitting. An assessment of whether the newly found decision boundaries are well representative for local conditions or not is however still necessary as otherwise inexperienced personnel might be misled by the results. We therefore recommend that the evaluation of the decision boundaries is done by geotechnicians who are generally experienced and familiar with the local conditions. In the presented case study, the assignment of core logs to the CPT data as well as the final evaluation of the new boundaries was done in such a way. The fact that the ANNs put ST 2 (i.e., sediments with organic material/peat) predominantly in an area of high $F_r$ but medium $Q_{tn}$ (as opposed to Robertson, 2009), where organic sediments should be in high $F_r$ and low $Q_{tn}$; see Figure 12 top row) indicates that these sediments contain a considerable amount of coarse grained material so that the tip resistance increases during the CPT test. Albeit this contradicts the Robertson charts, this is in good accordance with the original core logs and local experience, where it is observed that there is a considerable amount of sand and even gravel mixed within the organic material.

Beyond this, several other boundary conditions must be heeded to generate meaningful results: CPTs are usually performed as a part of geotechnical underground investigations to answer specific questions about the local soil stratigraphy. Considering this, different project sizes and the variable underground conditions in countries like Austria or Norway, a large, “real world” dataset can hardly be perfectly balanced. However, if a methodology like the one presented is to be used, the dataset upon which the data driven decision boundaries are based must sufficiently represent all classes. Figure 5 shows that the original dataset of this study is not well balanced, but we counteracted by using the SMOTE algorithm. By inspection of the newly generated datapoints and comparison to where the original datapoints were located, we presume that no additional bias was introduced by oversampling the underrepresented classes. Visualization of the data before and after preprocessing is recommended to guarantee that preprocessing worked as intended (e.g., Figure 7).

The procedure of using an ensemble of different ANNs with randomly generated architectures is also seen as beneficial, as enough randomness is introduced to achieve a wide range of unique classifications. However, the learned decision boundaries from the $Q_{tn}$ versus $F_r$ chart of the Zell basin (third row in Figure 12), which looks somewhat “patchy” indicates that there is still room for improvement. The generated classification where individual soil types show up multiple times in different places might lead to a higher classification accuracy for this subset of data, but it will also decrease the generalizability of this chart which in return indicates overfitting.

Additionally, to the above given boundary conditions, one should consider the strong dependence of the results on the assigned class labels. Due to the big overall size of the dataset of this study, the core logs that were used as labels have been mapped by many different geotechnicians. We see this as positive, as the labels reflect the mapping skills of many different experts and are not biased towards individuals. As core logging is still mostly a “manual” task (aided by laboratory tests), the distinction between ST 5 (i.e., clayey silt to fine-sandy silt) and ST 6 (i.e., clayey silt to clay—silt mixtures) is prone to errors. Manually differentiating between such fine grain sizes is barely possible and reflected in the present dataset by a high overlap between the datapoints of ST 5 and ST 6.

### 6 | OUTLOOK

The above given discussion shows that the proposed methodology is a new way to find locally adjusted soil type boundaries based on supervised machine learning. A practical application would be the implementation of learned decision boundaries that have been evaluated by locally experienced engineers in CPT data interpretation software. On the one hand this would aid inexperienced geotechnicians who are unfamiliar with the local ground conditions to correctly interpret new CPT tests, as they could only rely on general CPT interpretation charts before (see the comparison of the learned boundaries against the general decision boundaries in Section 4). On the other hand, decision boundaries that were learned from a sufficient amount of CPTs and core drillings would diminish the need for further, expensive core drillings as more information could be gathered from CPTs alone. Another benefit of the methodology is that once new decision boundaries are found, they can always be updated when new data (i.e., new CPT tests and new core logs) becomes available. The framework shall however not be used as a replacement, but as an additional aid for CPT data interpretation.

While future studies might on the one hand address improvements concerning the data preprocessing, using more advanced algorithms instead of MLPs should be considered (e.g., enhanced probabilistic neural networks (Ahmadlou & Adeli, 2010); Neural Dynamic Classification algorithm (Rafiei & Adeli, 2017); FEMa (Pereira, Piteri, Souza, Papa, & Adeli, 2020)). Albeit using ensembles of random ANN architectures showed good results, new methods for ensemble design like the Dynamic Ensemble...
Learning Algorithm (Alam, Siddique, & Adeli, 2020) are also worth consideration.

As the goal of this study was to find new boundaries between soil types, no efforts were undertaken to maximize the classification accuracy of individual CPT tests. However, training a classifier to directly “interpret” or classify data from a CPT test as accurately as possible is a worthwhile endeavor. Approaches to do so would be to use more than two input dimensions as input at once (e.g., use $Q_{tn}$, $F_r$, and $U_2$) which might help an ANN to find better class boundaries within the overlapping classes. Additionally, a classifier might be presented with “windows” of multiple datapoints at the same time instead of singular points to retrieve information from the local neighborhood. Upcoming studies will go into such a direction.

ACKNOWLEDGMENTS

Dr. Michael Premstaller is gratefully acknowledged for providing the main part of the dataset, as well as the Norwegian Geotechnical Institute which provided data from the NGTS study. Assistant Professor Franz Tschuchnigg is thanked for additional technical advice concerning CPTs.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Retrieved from https://www.tensorflow.org/

Ahmadlou, M., & Adeli, H. (2010). Enhanced probabilistic neural network with local decision circles: A robust classifier. Integrated Computer-Aided Engineering, 17(3), 197–210.

Alam, K. M. R., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. Neural Computing and Applications, 32(12), 8675–8690.

Bishop, C. M. (2006). Pattern recognition and machine learning. Information science and statistics. New York, NY: Springer.

Blaker, Ø., Carroll, R., Paniagua, P., DeGroot, D. J., & L’Heureux, J.-S. (2019). Halden research site: Geotechnical characterization of a post glacial silt. AIMS Geosciences, 5(2), 184–234.

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

Carvalho, L. O., & Ribeiro, D. B. (2019). Soil classification system from cone penetration test data applying distance-based machine learning algorithms. Soils and Rocks, 42(2), 167–178.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Cherkauer, K. J. (1996). Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. Working notes, integrating multiple learned models for improving and scaling machine learning algorithms workshop, Portland. Palo Alto, CA: AAAI Press.

Chollet, F. (Ed.) (2018). Deep learning with Python. Shelter Island, NY: Manning, Safari Tech Books Online.

Chollet, F. and others. (2015). Keras.
Pereira, D. R., Piteri, M. A., Souza, A. N., Papa, J. P., & Adeli, H. (2020). FEMA: A finite element machine for fast learning. Neural Computing and Applications, 32(10), 6393–6404.

Quinteros, S., Gundersen, A., L’Heureux, J.-S., Antonio, H., Carraro, J., & Jardine, R. (2019). Oysand research site: Geotechnical characterisation of deltaic sandy-silty. AIMS Geosciences, 5(4), 750–783.

Rafiei, M. H., & Adeli, H. (2017). A new neural dynamic classification algorithm. IEEE Transactions on Neural Networks and Learning Systems, 28(12), 3074–3083.

Raschka, S., & Mirjalili, V. (2019). Python machine learning—Third Edition: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Birmingham, UK: Packt Publishing.

Robertson, P. K. (1991). Soil classification using the cone penetration test: Reply. Canadian Geotechnical Journal, 28(1), 176–178.

Robertson, P. K. (2009). Interpretation of cone penetration tests — a unified approach. Canadian Geotechnical Journal, 46(11), 1337–1355.

Robertson, P. K. (2016). Cone penetration test (CPT)-based soil behaviour type (SBT) classification system — an update. Canadian Geotechnical Journal, 53(12), 1910–1927.

Rogiers, B., Mallants, D., Batelaan, O., Gedeon, M., Huysmans, M., & Dassargues, A. (2017). Model-based classification of CPT data and automated lithostratigraphic mapping for high-resolution characterization of a heterogeneous sedimentary aquifer. PLoS One, 12(5), e0176656.

Schneider, J. A., Hotstream, J. N., Mayne, P. W., & Randolph, M. F. (2012). Comparing CPTU Q – F and Q – Δu2 /σv0 ‘soil classification charts. Géotechnique Letters, 2(4), 209–215.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.

How to cite this article: Erharter GH, Oberhollenzer S, Fankhauser A, Marte R, Marcher T. Learning decision boundaries for cone penetration test classification. Comput Aided Civ Inf. 2021;36:489–503.

https://doi.org/10.1111/mice.12662