Title

A maching learning-based analysis of liquefaction input factors using the Next Generation Liquefaction database

Authors

Zimmaro, Paolo
Hudson, Kenneth
Ulmer, Kristin
et al.

Publication Date

2024-07-01

Peer reviewed

18TH WORLD CONFERENCE ON EARTHQUAKE ENGINEERING
WCEE2024
MILAN

MILAN, ITALY
30TH JUNE - 5TH JULY 2024

www.wcee2024.it

A MACHINE LEARNING-BASED ANALYSIS OF LIQUEFACTION INPUT FACTORS USING THE NEXT GENERATION LIQUEFACTION DATABASE

P. Zimmaro[1,2], K.S. Hudson[2], K.J. Ulmer[3], S.J. Brandenberg[2], J.P. Stewart[2], & S.L.Kramer[4]

[1] University of Calabria, Rende, paolo.zimmaro@unical.it

[2] University of California, Los Angeles, Los Angeles, USA

[3] Southwest Research Institute, San Antonio, USA

[4] University of Washington, Seattle, USA

Abstract: *Liquefaction triggering is typically predicted using fully-empirical and/or semi-empirical models. Hence, such models are heavily reliant upon available liquefaction (and/or lack thereof) case history data. These predictive models are based on a variety of factors, describing the demand (i.e., the cyclic stress ratio, CSR in existing legacy models) and the capacity (i.e., the cyclic resistance ratio, CRR). However, the degree to which these factors truly affect models' performance is unknown. To explore this aspect and quantitatively rank the importance of liquefaction input model parameters, we leverage a Random Forest Machine Learning (ML) approach using two methods: (1) a feature importance metric based on the Gini impurity index, and (2) a SHapley Additive exPlanations (SHAP)-based approach. Both approaches were employed using typical input factors used in legacy liquefaction triggering models based on cone penetration test (CPT) data. Such analyses were performed using all reviewed (i.e., fully vetted) data in the Next Generation Liquefaction (NGL) database. Our analysis then separately explores the impact on resulting models of seven input parameters. We show that the most important input parameters are: (1) the peak ground acceleration, (2) the soil behavior type index, and (3) the earthquake magnitude (which serves as a proxy for duration in such models). The input parameters with the lowest importance are the total and the effective vertical stresses. A limitation of this analysis is that the ML model used does not allow for extrapolation beyond the range of the data. As a result, for input parameters with narrow distributions of the data (i.e., somewhat limited parameter space), a lower ranking could be associated with such limited availability of a wide range of values, rather than being related to actual low importance. This limitation likely accounts for the low importance attached to stress-related input parameters since legacy case histories are generally related to shallower (<10m) depths.*

1 Introduction

Soil liquefaction is one the most damaging geotechnical hazards triggered by strong earthquakes. As a result, prediction models are very important to identify sites where liquefaction may occur during future earthquake events. Current liquefaction triggering models typically rely upon databases of liquefaction performance at a site (i.e., the presence or absence of surface manifestations of liquefaction), measured in-situ penetration resistance by means of standard penetration tests (SPT; e.g., Boulanger and Idriss, 2012; Cetin et al., 2004 and 2018) and/or cone penetration tests (CPT; e.g., Boulanger and Idriss, 2016; Moss et al., 2006), and recorded or estimated ground motion levels at the site. These legacy models are typically empirical or semi-empirical. Hence they are heavily reliant on quantity and quality of data. Since their development, many,

potentially hundreds, of new liquefaction case histories were collected. As a result, this new data may become consequential when revisiting such models, potentially informing novel and innovative methods.

The Next Generation Liquefaction (NGL) project aims at bridging this knowledge gap by creating an open source relational database (Brandenberg et al. 2019; Zimmaro et al., 2019; Ulmer et al., 2023) that contains all legacy case histories and a large number of case histories from more recent earthquakes including the Canterbury (New Zealand) 2010-2011 earthquake sequence, the Tohoku (Japan) 2011 earthquake, the Emilia (Italy) 2012 earthquake sequence, and the 2016 Kaikoura (New Zealand) 2016 event.

In recent years the increasing availability of high quality liquefaction case history data enabled new big-data-based approaches, including liquefaction prediction models using artificial intelligence (AI)-based models (e.g., Durante and Rathje, 2021; Maurer and Sanger, 2023 and references therein). The NGL database offers a unique opportunity to train and develop such models. As a result, we trained multiple random forest (RF) machine learning (ML) models using CPT data and different permutations of input parameters derived using a critical layer-based approach (i.e., for each CPT profile, we selected a single layer responsible for the observed site performance – Yes/No liquefaction manifestations). The resulting underlying model was then leveraged to identify the most important input parameters, i.e., those that most strongly influence the final outcome of these models. This is relevant as outcomes from this study may inform future model development efforts.

2 Data and Methods

2.1 Database of liquefaction case histories

The data used in this study were retrieved from the NGL database on July 10, 2023. The query to extract such data was performed on reviewed data only (i.e., data fully vetted following standard protocols developed by the NGL database working group, Zimmaro et al., 2019). Only case histories with available CPT data were used in this analysis. The CPT data were interpreted within a critical layer-based approach – for each CPT profile, data were extracted for a single layer, considered to be the one most likely to have produced the liquefaction observations made at the site or, in the cases where surface manifestation was not observed to have been the layer most likely to have produced surface manifestation if the shaking had been strong enough to do so. CPT layers in this analysis were derived using the approach proposed by Hudson et al. (2023). The resulting database consists of a total of 546 liquefaction manifestation data points (191 No; 355 Yes). The following factors were considered in the analysis: (1) earthquake moment magnitude (**M**), (2) recorded or estimated peak ground acceleration at the site (PGA), (3) depth to ground water (GWT), (4 and 5) layer total and effective vertical stress (σ_v and σ'_v, respectively), (6) the CPT equivalent clean sand correction of the normalized tip resistance ($q_{c1N,cs}$), and (7) the CPT-based soil behavior type index (I_C, Robertson, 1990).

Figure 1 shows histograms illustrating the parameter space covered by these input factors. Four different combinations of the seven selected input parameters (Table 1) were used to evaluate the stability of the final parameter rankings with respect to different input permutations.

2.2 Machine learning and importance metrics

Several ML approaches may be used to develop liquefaction prediction models importance ranking leveraging the NGL-based dataset presented in the previous sub-section. In this study we selected the Random Forest (RF) model (Breiman, 2001) to train the data, test candidate models, develop a final prediction model, and extract feature importance rankings. A schematic of a RF model applied to a binary problem with one decision tree and two layers (i.e., depth = 2) is shown in Figure 2. RF algorithms are tree-based approaches. They are based on many decision trees combining different input features. When developing the model, the prediction space is subdivided into smaller regions (encompassing the entire parameter space) using simple Yes/No questions. Each of these smaller regions constitute a decision tree. Each decision tree leads to a final solution. In binary problems such as that analyzed in this study (i.e., Yes/No with respect to the observation of liquefaction manifestations), the solution can only be Yes or No. Once all decision trees are built (i.e., the "random" forest is created), the answer with the most votes (where a vote consists in a final outcome of a single decision tree; i.e., the outcome appearing in the majority of leaves) in the whole RF is the final solution. This model can capture relationships between features in the dataset and it is easy to interpret (i.e., the algorithm process can be visualized and explained, avoiding black boxes). Each decision in each tree leads to a split that has a probability of misclassifying an observation. This probability can be calculated as 1 - *p(i)*

where *p(i)* is the fraction of observations correctly predicted (e.g., the observation is Yes and the outcome of the split is also Yes). In Figure 2, a total of three splits are reported. The first split consists in subdividing all samples (i.e., each observation data point, 10 in Figure 2) into predictions characterized by PGA values greater or smaller than 0.25. Two more splits occur at the second layer of the decision tree and involve decisions pertaining to **M** and I_C. Misclassifications can be quantified and calculated using the Gini impurity index (Breiman, 1984):

$$\text{Gini impurity index} = \sum_{i=1}^{C} p(i) \cdot [(1 - p(i)] \quad (1)$$

where *C* is the total number of classes (2 in binary problems) and *p(i)* is the probability that the i^{th} data point falls in that class. If the Gini impurity index is equal to zero, the split did not misclassify any data point. The lower this value, the better the split within the process. In Figure 2, Gini impurity indexes associated to each split are shown.

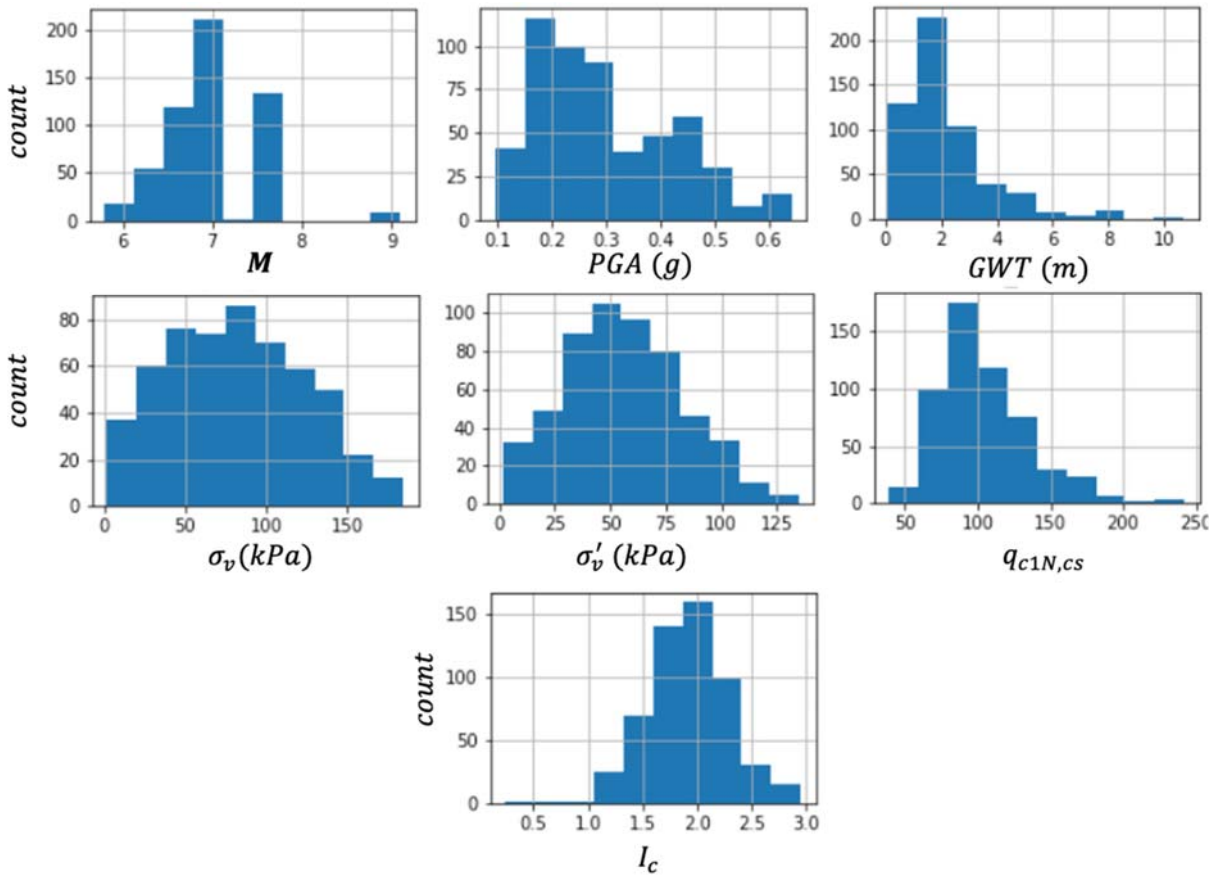

Figure 1. Distribution of all input parameters used in the analysis.

Table 1. Overview of the input parameters used in each RF set.

Parameter	RF Set #1	RF Set #2	RF Set #3	RF Set #4
σ_v (kPa)	✓	✓	X	X
σ'_v (kPa)	X	X	✓	✓
PGA (g)	✓	✓	✓	✓
M	✓	✓	✓	✓
$q_{c1N,cs}$	✓	✓	✓	✓
I_c	X	✓	X	✓
Depth to GW (m)	✓	✓	X	X

When building a RF model, the dataset should be subdivided into training and testing subsets. The model will then be trained on the training dataset and tested on the testing dataset (for which the outcome is unknown). For each of these phases, the model can be optimized using a performance metric (i.e., a metric that defines the reliability of the model based on the model's outcomes). In this study we adopt four different performance metrics. For each, we build a separate model, for each of the four combinations of the input parameters utilized. This leads to a total of 16 models and feature importance rankings. The following performance metrics were used:

(1) Recall (Eq. 2):

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

(2) Accuracy (Eq. 3):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

(3) Receiving Operator Curve (ROC) area under the curve; and
(4) Cohen's k coefficient (Eq. 4):

$$k = \frac{2 \cdot (TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)} \tag{4}$$

where TP and FP mean true positive and true negative, respectively (i.e., Yes and No manifestations where observed and correctly predicted), FP and FN stand for false positive and false negative, respectively (i.e., Yes and No manifestations were observed but mispredicted).

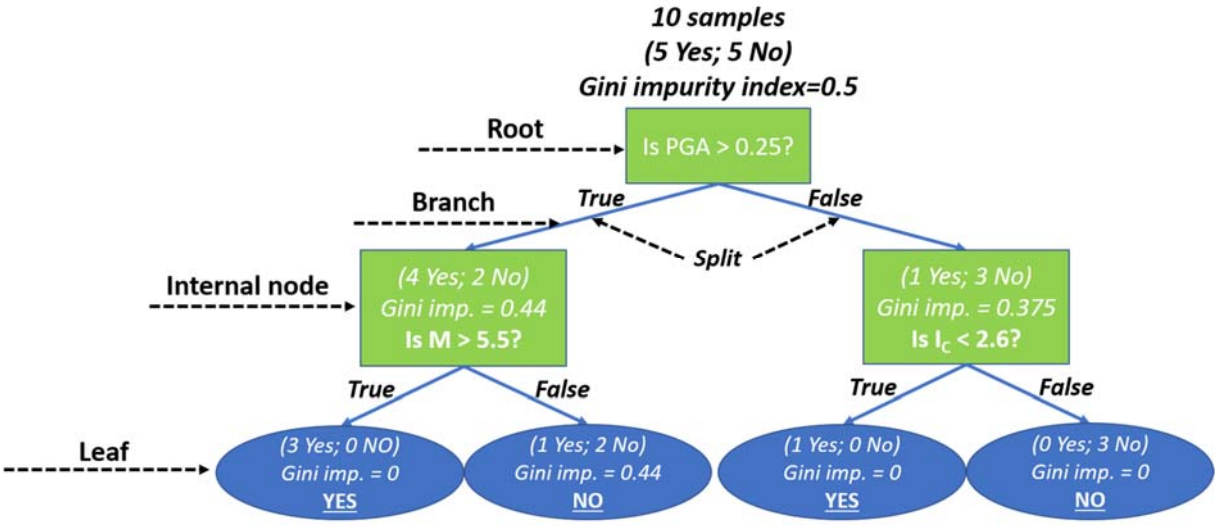

Figure 2. Schematic showing an example RF for a single decision tree for a binary problem.

In this study the training dataset comprises 80% of the data, while the testing dataset consists of the remaining 20% of the data. In addition to splitting the dataset, we also adopt a cross-validation approach in the training phase. This procedure randomly subdivides the dataset into (k) subsets, uses (k − 1) sets to train the model, and one set to validate it. In this study, k is set equal to 10. Tree-based approaches can easily overfit the data, resulting in high performance metrics in the testing dataset and lower performance metrics for the testing dataset. To avoid model overfitting, we carefully identified the number of trees and model set up values, to ensure that a similar performance is obtained for both the training and the testing phases.

Armed with the above-described 16 RF models, we evaluated two input parameter importance rankings for each model. These rankings are based on the following methods: (1) a Gini impurity index-based feature importance approach, and (2) a method based on the SHapley Additive exPlanations (SHAP) approach. The former is performed feature by feature and is based on how much, on average, the introduction of a feature decreases the Gini impurity index. The average over all trees in the forest is the measure of the feature

importance. Results are presented as a bar chart where each bar represents the feature importance of each input parameter. The latter is performed for each value assumed by each feature. The resulting SHAP ranking is then calculated based on how much the solution changes when each individual data point is not used to predict the outcome. Results are presented by means of a chart showing a number of columns equal to the number of features. Then, for each line (feature), all data points are shown, with an x-axis value representing how much the outcome changes when each individual data point is not used in the analysis. Values on this plot are color-coded based on the feature value. Changes on the x-axis can be positive or negative.

3 Results

In this section we focus on evaluating feature importance rankings for all developed models. Subsequent research will be focusing on the underlying RF models. These models all have similar performance metrics values, with values ranging between 0.7 and 0.76. These ranges apply to both the training and testing datasets, ensuring that the models developed in this study do not overfit the data. Since the database is unbalanced (the number of Yes cases is substantially higher than that of No cases), the depth of a decision tree and the number of trees were kept to relatively low numbers. In particular, we set the maximum depth (i.e., number of splits for each decision tree) to five, and the maximum number of trees in each forest (i.e., number of estimators) to 100.

Table 2 shows the first three features for each RF set and each performance metric. Both approaches (i.e., the Gini impurity index and SHAP methods) produce similar results. When they produce different outcomes, the relative feature for each method are shown. When both methods produce the same position in the ranking, only one feature is reported. As shown in Table 2, PGA is almost ubiquitously the most important feature in the ranking. When I_C was used, it is almost always the second highest ranked feature. **M** and $q_{c1N,cs}$ overall are the third and fourth most important features. Interestingly the level of stress usually ranks low in the ranking. We speculate this may be due to a relatively small range of variation of this parameter, which lacks entries for high values, corresponding to deeper layers.

Table 2. Summary of top three features by importance for all analysed models.

Model	Performance metrics	#1 Feature	#2 Feature	#3 Feature
RF Set #1	Accuracy	PGA	$q_{c1N,cs}$	Depth to GW
	Recall	PGA	**M**	$q_{c1N,cs}$
	ROC	PGA	**M** (SHAP), $q_{c1N,cs}$ (FI)	Depth to GW
	Cohen's k	PGA	Depth to GW	$q_{c1N,cs}$ (SHAP), σ_v (FI)
RF Set #2	Accuracy	I_c (SHAP), PGA (FI)	PGA (SHAP), I_c (FI)	Depth to GW
	Recall	PGA	I_c	**M** (SHAP), $q_{c1N,cs}$ (FI)
	ROC	PGA	I_c	**M** (SHAP), $q_{c1N,cs}$ (FI)
	Cohen's k	PGA	I_c	**M** (SHAP), Depth to GW (FI)
RF Set #3	Accuracy	PGA	σ'_v (SHAP), $q_{c1N,cs}$ (FI)	**M** (SHAP), σ'_v (FI)
	Recall	PGA	**M**	$q_{c1N,cs}$
	ROC	PGA	**M**	$q_{c1N,cs}$ (SHAP), σ'_v (FI)
	Cohen's k	**M** (SHAP), PGA (FI)	PGA (SHAP), **M** (FI)	$q_{c1N,cs}$
RF Set #4	Accuracy	PGA	I_c	**M** (SHAP) $q_{c1N,cs}$ (FI)
	Recall	PGA	I_c	**M**
	ROC	PGA	I_c	**M** (SHAP) $q_{c1N,cs}$ (FI)

	Cohen's k	PGA	I_c	**M**

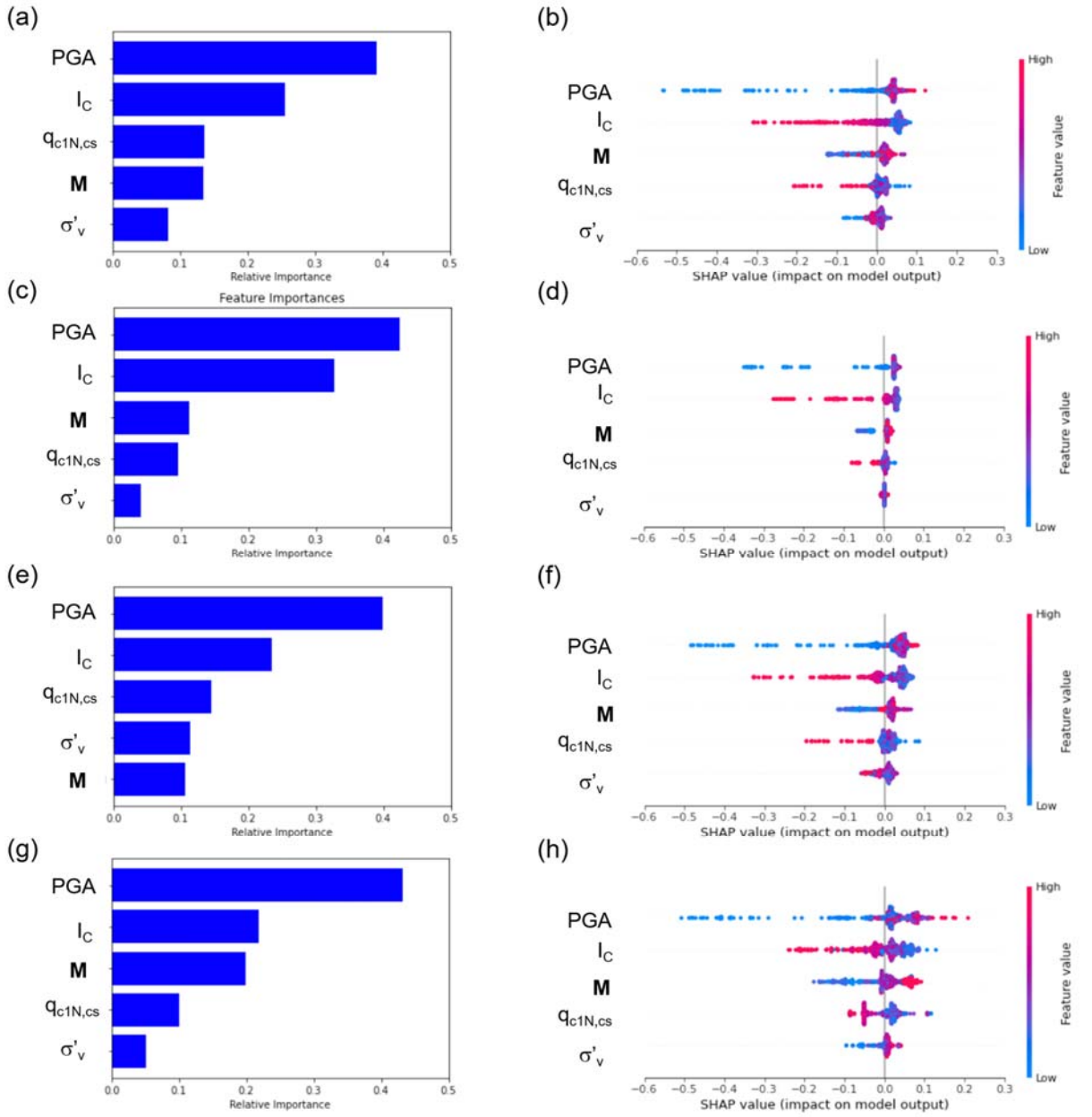

Figure 3. Relative importance of all analyzed features in RF #4 using the following performance metrics and importance ranking approaches: (a) accuracy, Gini impurity index-based feature importance, (b) accuracy, SHAP, (c) recall, Gini impurity index-based feature importance, (d) recall, SHAP, (e) ROC, Gini impurity index-based feature importance, (f) ROC, SHAP, (g) Cohen's k coefficient, Gini impurity index-based feature importance, and (h) Cohen's k coefficient, SHAP.

To better visualize the outcomes provided in Table, 2 we report, in Figure 3, for all performance metrics, the outcomes for the Gini impurity index-based feature importance and the SHAP method, for RF Set #4. Figure 3 confirms that PGA is always the most important feature and I_C is ubiquitously the second highest-ranked feature. Bar charts from the Gini impurity index-based feature importance analysis not only show that PGA is the highest ranked feature, but also that its importance is always significantly higher than all other features. The analysis of the SHAP method outcomes illustrates that feature value color codings follow specific trends,

which is a desirable feature from ML models. Low values of PGA are linked to negative SHAP values. This means that as PGA decreases, the model moves towards No predictions. Similar trends can be observed for all features. For instance, as I_C and $q_{c1N,cs}$ decrease, corresponding SHAP value increase, which means that as soil resistance decreases, it is more likely to observe liquefaction manifestations, and as the soil behavior type index goes towards clayey materials, manifestations are less likely to occur. It should be noted that despite I_C and $q_{c1N,cs}$ being highly correlated, our analyses show that the former is always more important than the latter when both factors are included in the prediction model (i.e., models 2 and 4). Also the trend of SHAP values with **M** is consistent with what we expected. As **M** increases (and so does the duration of shaking and the number of cycles – **M** is included in liquefaction prediction model as a proxy for duration), soil is expected to be more likely to produce liquefaction manifestations. The only exception to the meaningful trends observed in Figure 3 is represented by σ'_v. This input parameter has the lowest feature importance and a somewhat narrow and inexplicable trend. In Figures 3b,d,f it does not seem to show a trend at all. In Figure 3h it has a trend opposite to what was expected, with higher stresses (reflecting greater depths) corresponding to more likely-to-manifest layers. As mentioned earlier, a wider parameter space for this feature would certainly lead to more meaningful results and improved models.

4 Conclusions

We presented a ML-based analysis of the importance of various input features on liquefaction manifestation. The dataset presented in this study is a recent snapshot of vetted data available in the NGL database at a particular point in time. Such data was extracted for 546 case histories (191 No; 355 Yes). Each data point represents a single critical layer of fully-vetted CPT profiles. We developed a total of 16 RF models based on four model performance metrics (accuracy, recall, ROC area under the curve, and Cohen's k coefficient) and four different permutations of the seven selected input features: (1) **M**, (2) PGA, (3) depth to GW, (4) σ_v (5) σ'_v (6) $q_{c1N,cs}$, and (7) I_C. Leveraging these models we generated 32 feature importance rankings, using two methods: (1) an approach based on the Gini impurity index, and (2) a SHAP-based method. There is little between-method differences, highlighting that the two methods provide similar rankings for this problem and for the selected input features. We show that, overall, PGA is the most important feature. The second highest-ranked feature (when present in the developed models) is I_C, followed by **M** and $q_{c1N,cs}$. Surprisingly, when I_C is not included in the model (i.e., models 1 and 3), $q_{c1N,cs}$ has an overall importance that is usually lower than that of **M**. This is counterintuive as the strong correlation between I_C and $q_{c1N,cs}$ would suggest the opposite. This can be technically explained by the fact that I_C produced better splits in the resulting models (i.e., its inclusion in the models produces a stronger reduction in the Gini impurity index than $q_{c1N,cs}$). All features follow expected trends in the developed models, with the exception of total and effective stress values, which also have low-to-negligible importance. We speculate that this may be due to the fact that most stress values are related to relatively shallow layers and that their resulting parameter space is too narrow. We anticipate that the framework and the outcomes presented in this study would be useful for future development of liquefaction manifestation/triggering empirical and/or semi-empirical models and funding prioritization.

5 Acknowledgements

Financial support for the NGL project is provided by the U.S. Nuclear Regulatory Commission (NRC) and the U.S. Bureau of Reclamation (USBR) through the Southwest Research Institute (SWRI). Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product, or process disclosed in this paper, or represents that its use by such third party would not infringe privately owned rights. The views expressed in this paper are not necessarily those of the NRC or USBR.

6 References

Boulanger R.W., Idriss I.M. (2012) Probabilistic standard penetration test-based liquefactiontriggering procedure. Journal of Geotechnical & Geoenvironmental Engineering, 138: 1185–1195.

Boulanger RW and Idriss IM (2016) CPT-based liquefaction triggering procedure. Journal of Geotechnical & Geoenvironmental Engineering, 142(2): 04015065.

Brandenberg S.J., Zimmaro P., Stewart J.P., Kwak D.Y., Franke K.W., Moss R.E.S., Cetin K.O., Can G., Ilgac

M., Stamatakos J., Weaver T., Kramer S.L. (2019). Next generation liquefaction database. Earthquake Spectra, 36, 939-959.

Breiman L. (1984). Classification and Regression Trees (first edition). Routledge, London (United Kingdom). DOI: 10.1201/9781315139470.

Breiman L. (2001) Random forests. Mach. Learn., 45, 5–32.

Cetin K.O., Seed R.B., Der Kiureghian A., Tokimatsu K., Harder Jr L.F. Kayen R.E., Moss R.E.S. (2004). SPT-Based probabilistic and deterministic assessment of seismic soil liquefaction potential. Journal of Geotechnical & Geoenvironmental Engineering, 130(12): 1314–1340.

Cetin K.O., Seed RB., Kayen R.E., Moss R.E.S., Bilge H.T., Ilgac M., Chowdhury K. (2018). SPT-based probabilistic and deterministic assessment of seismic soil liquefaction triggering hazard. Soil Dynamics and Earthquake Engineering, 115:698-709.

Durante M.G., Rathje E.M. (2021). An exploration of the use of machine learning to predict lateral spreading. Earthquake Spectra, 37(4): 2288-2314.

Hudson K.S., Ulmer K.J., Zimmaro P., Kramer S.L., Stewart J.P., Brandenberg S.J. (2023). Unsupervised machine learning for detecting soil layer boundaries from cone penetration test data. Earthquake Engineering and Structural Dynamics. DOI: 10.1002/eqe.3961.

Maurer B.W., Sanger M.D. (2023). Why “AI” models for predicting soil liquefaction have been ignored, plus some that shouldn’t be. Earthquake Spectra, 39(3):1883-1910.

Moss R.E.S., Seed R.B., Kayen R.E., Stewart J.P., Der Kiureghian A., Cetin K.O. (2006). CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. Journal of Geotechnical & Geoenvironmental Engineering, 132(8):1032-1051.

Robertson P.K. (1990). Soil classification using the cone penetration test. Can Geotech J., 27(1):151-158.

Ulmer K.J., Zimmaro P., Brandenberg S.J., Stewart J.P., Hudson K.S., Stuedlein A.W., Jana A., Dadashiserej A., Kramer S.L., Cetin K.O., Can G., Ilgac M., Franke K.W., Moss R.E.S., Bartlett S.F., Hosseinali M., Dacayanan H., Kwak D.Y., Stamatakos J., Mukherjee J., Salman U., Ybarra S., Weaver T. (2023). Next-Generation Liquefaction Database, Version 2. Next-Generation Liquefaction Consortium. DOI: 10.21222/C23P70.

Zimmaro P., Brandenberg S.J., Bozorgnia Y., Stewart J.P., Kwak D.Y., Cetin K.O., Can G., Ilgac M., Franke K.W., Moss R.E.S., Kramer S.L., Stamatakos J., Juckett M., Weaver T. (2019). Quality control for next-generation liquefaction case histories. 7th International Earthquake Geotechnical Engineering for Protection and Development of Environment and Constructions- Proceedings of the 7th International Conference on Earthquake Geotechnical Engineering, VII ICEGE, pp. 5905 – 5912, Rome (Italy), June 17-20.

Zimmaro P., Brandenberg S.J., Stewart J.P., Kwak D.Y., Franke K.W., Moss R.E.S., Cetin K.O., Can G., Ilgac M., Stamatakos J., Juckett M., Mukherjee J., Murphy Z., Ybarra S., Weaver T., Bozorgnia Y., Kramer S.L. (2019). Next-Generation Liquefaction Database. Next-Generation Liquefaction Consortium. DOI: 10.21222/C2J040.