Earthquake-Induced Liquefaction Manifestation Multiclass Prediction Utilizing Random Forests for the Canterbury Earthquake Sequence

Katherine Cheng, S.M.ASCE[1]; Pablo Busch[2]; and Katerina Ziotopoulou, Ph.D., P.E., M.ASCE[3]

[1]Ph.D. Candidate, Dept. of Civil and Environmental Engineering, Univ. of California, Davis, Davis, CA. Email: katcheng@ucdavis.edu
[2]Ph.D. Student, Energy and Efficiency Institute, Univ. of California, Davis, Davis, CA. Email: pmbusch@ucdavis.edu
[3]Associate Professor, Dept. of Civil and Environmental Engineering, Univ. of California, Davis, Davis, CA. Email: kziotopoulou@ucdavis.edu

ABSTRACT

The abundance of post-earthquake data from the Canterbury, New Zealand (NZ), area can be leveraged for exploring machine learning (ML) opportunities for geotechnical earthquake engineering. Herein, random forest (RF) is chosen as the ML model to be utilized as it is a powerful non-parametric classification model that can also calculate global feature importance post-model building. The results and procedure are presented of building a multiclass liquefaction manifestation classification RF model with features engineered to preserve special relationships. The RF model hyperparameters are optimized with a two-step fivefold cross-validation grid search to avoid overfitting. The overall model accuracy is 96% over six ordinal categories predicting over the Canterbury earthquake sequence measurements from 2010, 2011, and 2016. The resultant RF model can serve as a blueprint for incorporation of other sources of physical data such as geological maps to widen the bounds of model usability.

INTRODUCTION

While the fundamental understanding of liquefaction mechanics at the element level has been established, it is a nonlinear phenomenon with a wide range of potential participating input parameters that often have interactive effects particularly at the system level of a soil deposit. A possible manifestation of liquefaction is the formation of ejecta or sand boils at the ground surface by the seepage of water through ground cracks and the simultaneous temporary loss of the soil's bearing capacity. Liquefaction manifestation is a multiscale problem with many interacting factors that are challenging to isolate. For example, even if someone studies the response of a singular sand via an exhaustive direct simple shear (DSS) testing program, upscaling to the field is not straightforward since a liquefiable soil site operates as a system of multiple layers and/or inclusions that may affect the observed manifestations (e.g., Cubrinovski et al. 2019, Bassal et al. 2022). As such, case histories become invaluable one-off opportunities towards studying multiple aspects of liquefaction.

Combined with advancements in field instrumentation, monitoring, data storage, and analysis capabilities, case histories have yielded rich and high quality during- and post-earthquake data. These newer datasets allow for liquefaction phenomena to be explored with more advanced techniques, such as machine learning (ML). A subset of Artificial Intelligence (AI), ML uses computer algorithms to make predictions on given data without explicit programming instructions. A ML algorithm can find patterns and relationships among the data without explicit guidance and could potentially discover associations that may otherwise be missed.

New Zealand is a seismically active island country and has experienced several major earthquakes. The 2010/2011 Canterbury earthquake sequence (CES) started with the M_w7.1 Darfield earthquake (September 4, 2010) and was followed by up to ten events which induced liquefaction in the affected region (e.g., Maurer et al. 2015). The soil deposits in Christchurch, a city located in the Canterbury region, vary significantly both horizontally and vertically and are composed of a mix of thin layers of sand, silt, clay, and peat (Cubrinovski et al. 2011, Beyzaei et al. 2018). The depositional history of the soils was unable to be captured solely by traditional CPT-based liquefaction triggering assessments which resulted in numerous false positive or false negative predictions of liquefaction manifestation occurrences (Ballegooy et al. 2014, Beyzaei et al. 2018). The combination of available information on these complex deposits with well documented liquefaction manifestation data yielded a seminal case history. This data has been used in ML applications for earthquake engineering, some involving Random Forest (RF) for binary liquefaction manifestation occurrence classification (Durante and Rathje 2021) or for liquefaction potential index (LPI) prediction (Geyin et al. 2022).

Herein, the application of the RF ML algorithm on an established large dataset is investigated. In the process the methods and approaches behind a successful application of ML in geotechnical earthquake engineering are demonstrated and established. The work presented uses post-processed data from the September 2010, February 2011, and February 2016 Canterbury, New Zealand earthquakes (Geyin et al. 2021) to predict different severities of liquefaction manifestation in the form of multiple categories via the Random Forest ML algorithm (multiclass classification). The paper first introduces the dataset and RF algorithm with a subsection justifying the choice of algorithm. Then the subsequent feature engineering on the dataset is explained along with details of the two-stage cross-validation used to optimize the RF hyperparameters while minimizing overfitting. Finally, the fitted RF model is presented along with the final accuracies and model performance results based on various evaluation metrics. The paper concludes with lessons learned from the application of RF with this dataset, the feasibility of incorporating physical characteristics with ML in the field of liquefaction, and challenges and opportunities for the future with incorporation of physical relationships.

DATASET

The dataset is composed of earthquake and CPT data from the September 2010, February 2011, and February 2016 earthquakes in Canterbury, New Zealand (Geyin et al. 2021). The number of CPT observations is 5,668. It is a structured dataset containing typical CPT data, recorded peak ground acceleration (PGA) in *g* and groundwater table (GWT) depth in *m* for each CPT location per earthquake, and the severity of liquefaction manifestation experienced at each CPT per earthquake. The liquefaction manifestation category is the prediction target of the RF model. It can take the following values: 0 (no manifestation), 1 (minor), 2 (moderate), 3 (severe), 4 (lateral spreading), 5 (severe lateral spreading), and 10 (unknown). Further details of the separation of categories can be found in Geyin et al. (2021). All records corresponding to manifestation category 10 were removed from the dataset as they are essentially unknown values that took up only 5% of the total dataset. Only the 4 m of soil below the GWT were considered as that is the depth of soil with the most impact on manifestations during the New Zealand earthquakes (Cubrinovski and Robinson, 2016). The GWT fluctuates per year and the PGA is different for each earthquake, therefore a different set of data for the same CPTs is created for each earthquake year due to the subsequent feature engineering, roughly tripling the base dataset.

RANDOM FOREST ALGORITHM

The method chosen to predict liquefaction manifestation categories is the RF algorithm. The RF algorithm is an extension of the decision tree algorithm, in that it fits multiple decision trees and uses a majority vote approach between the trained decision trees to make its predictions as shown in Figure 1. The overall goal of the decision tree algorithm is to use the supplied predictor variables and recursively divide the dataset into homogeneous groups to accurately predict the response variable. In the fitting process of the RF algorithm, several hyperparameters are specified before training, such as the number of decision trees. Sampling with replacement (bootstrapping) at a certain percentage of the training set size is used to create new training datasets that are assigned to each of the individual decision trees, with each training set being unique from one another due to the bootstrapping method. These bootstrapped datasets contain only a subset of the overall predictors in the training set both from the size and replacement sampling restrictions. From there, the decision trees each begin a recursive split process on their assigned datasets to create a set of rules that determine a given observation's classification. Each decision tree path to a split is called a branch, and where a branch splits is called a node, with a terminal node called a leaf node. In the training process for each decision tree, the algorithm first determines an optimal variable to split upon, and then an optimal value within the variable to separate the data. This process repeats recursively until various stop criteria have been reached. Some of the stop criteria for a decision tree are:

- **Minimum split size**: A threshold for how small the number of observations within each subgroup of data should be to continue splitting.
- **Minimum node size**: A threshold for how small the number of observations in a node should be to terminate the branch and convert the node into a leaf node.
- **Tree depth**: The number of recursive splits or nodes used to build the tree.

Once the controlling stop criterion is reached for each decision tree, each of the final splits within the decision trees are determined. For a new data point (observation), the RF algorithm checks the resulting classification from each decision tree in its "forest" and then uses a majority vote approach to predict the observation's final classification. Besides Scikit-learn (Pedregosa et al. 2011) for the RF algorithm, other packages used to prepare the data and process the predictions include NumPy (Harris et al., 2020), Pandas (McKinney, 2010), and Yellowbrick (Bengfort et al. 2018).

Here, the choice was made to use only one ML model, RF, while focusing on feature engineering with an emphasis on physical ties and overfitting mitigation such that it can provide insights that will enhance future adoption into a more generalized usage. RF was chosen as the final model because: (i) Variable interactions are preserved in the RF algorithm process which is pertinent for liquefaction analysis; ii) RF is a non-parametric method as it has no prior data distribution (e.g., Normal) requirements for the algorithm and is known to be able to fit nonlinear parameter relationships (Brieman, 2001); (iii) RF yields low bias with only moderate variance which indicates that the predictions will tend to center around the correct category; (iv) RF models carry a degree of interpretability as they present the feature importance of the explanatory variables included. Feature importance allows one to draw the results back to existing knowledge of liquefaction phenomena behavior and check if there is agreement between the most influential features. If there are any unexpected results, they can serve as an indication that the model may not be performing within known physical bounds.

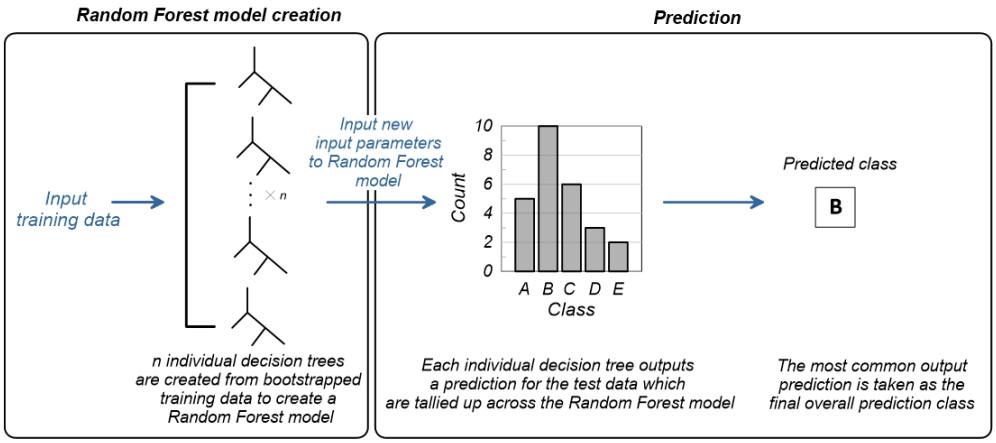

Figure 1. Conceptual Random Forest (RF) algorithm process

FEATURE ENGINEERING

Feature engineering is the practice of extracting features from raw data (such as CPT data) with the help of domain knowledge. Following Cubrinovski and Robinson's (2016) observation that the 4 m of soil below the GWT had the most impact on liquefaction manifestation for New Zealand earthquakes, each CPT in the dataset was first truncated to only contain the data from the GWT for the given year to the subsequent 4 m below it. The GWT depth per CPT was also an input feature, as it contains the elevation and saturation information per CPT. After truncation, the CPT section was split via soil behavior type (SBT) into a set of soil layers. SBT was calculated via Robertson's (2010) cutoffs for each soil type with an assumption of 18 kN/m^3 for the saturated soil unit weight. The SBT categories consist of six types of soils: organic soils, clay silt mixtures, sand mixtures, sands, and gravelly to dense sands. The distance between each CPT reading is 0.02 m and a soil layer consists of one continuous SBT section within the same CPT. The CPT is then discretized into subsections of soil layers and the following data is recorded for each layer thus creating the feature space: depth at the top of the soil layer [m], GWT [m], thickness of the soil layer [m], type of soil based on SBT, peak ground acceleration (PGA) [g], category of liquefaction manifestation, and location label.

The location label feature consists of a spatial label that indicates the geographical location in terms of Northings and Eastings. A 0.01° (1.11 km) distance was used as the grid distance in Northing and Easting directions to separate each location section geographically. Any CPTs within the same grid section received the same numerical label, with the extent and density of the grid shown in Figure 2. The location labels were created by assigning an integer to each grid square of size 0.01° in order, starting from the top left corner and moving along the row to the right. After a row is labeled, the algorithm continues labeling from left to right on the subsequent row until the last row.

As the PGA and GWT differ for each year for each location, each earthquake event yielded a different set of data for the same CPT locations. Figure 3 illustrates the counts of liquefaction manifestation categories per earthquake event, labeled by year of earthquake occurrence, showcasing the imbalanced dataset. After this process of feature engineering the original dataset with all three earthquakes, the new dataset contained 239,122 observations. The features were all preserved in their original units and no feature transformation was used as the intention was to

keep consistent physical units throughout the whole process. Table 1 contains the summary statistics and distributions of both the input features and target values used in the model.

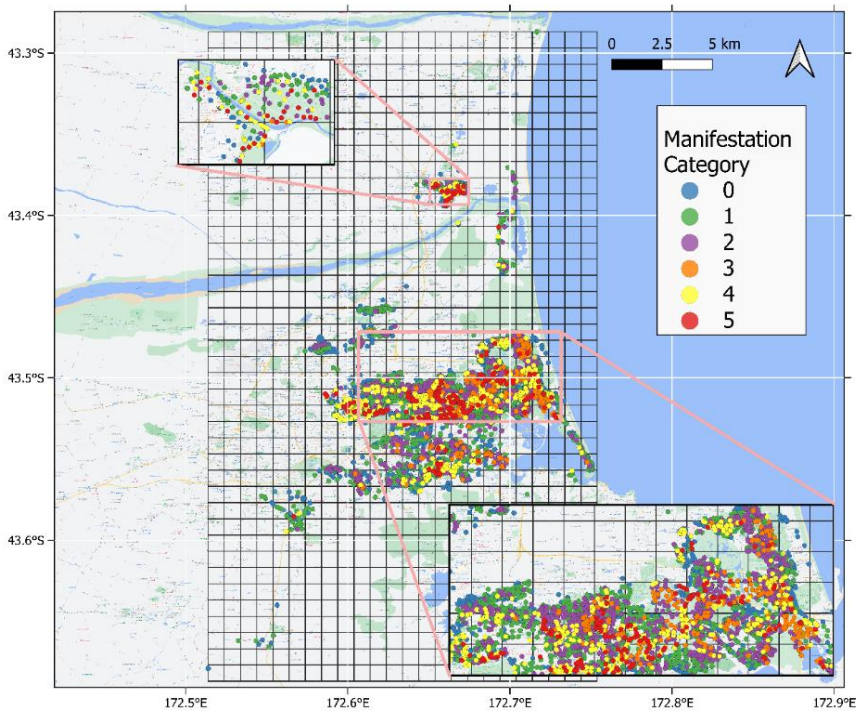

Figure 2. CPT observations in the study area from Canterbury, New Zealand. ***Source: Geyin et al. (2021).***

Table 1. Summary statistics for variables

Variable	Count	Mean	Std	Min	Median	Max	Earthquake year	Count	%total
Depth (m)	239,122	3.41	1.35	0	3.33	10.5	*2010*	80,024	33.50%
Length (m)	239,122	0.24	0.45	0.01	0.1	3.99	*2011*	82,676	34.60%
PGA (g)	239,122	0.25	0.13	0.05	0.21	0.73	*2016*	76,422	31.90%
GWT (m)	293,122	1.69	0.76	0	1.6	6.78			

Manifestation level	Count	%total	SBT – Soil Type	Count	%total
0	147,445	61.70%	*2 - Organic Soils*	1,709	0.70%
1	43,748	18.30%	*3 - Clay*	18,459	7.70%
2	26,453	11.10%	*4 - Silt mixtures*	43,919	18.40%
3	5,600	2.30%	*5 - Sand mixtures*	61,695	25.80%
4	11,555	4.80%	*6 - Sands*	75,984	31.80%
5	4,321	1.80%	*7 - Gravelly to Dense sands*	37,356	15.60%

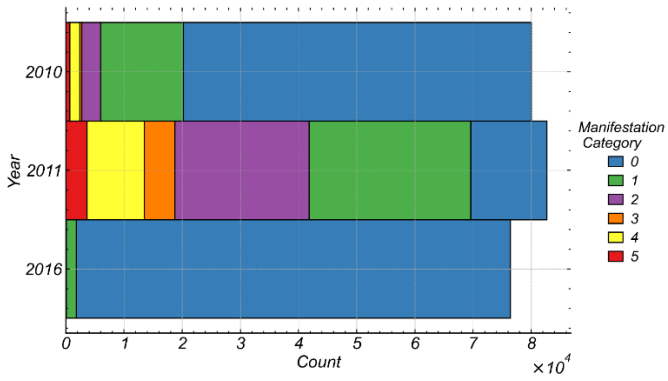

Figure 3. Count of manifestation events for the September 2010, February 2011, and February 2016 earthquakes in Canterbury, New Zealand

HYPERPARAMETER OPTIMIZATION

The dataset was split into two portions with 80% of it for training and 20% of it for testing. To improve the performance of the RF algorithm the hyperparameters were optimized via a random hyperparameter grid search with cross-validation on the training set. The global grid search consisted of the hyperparameters in Table 2, resulting in a total of 1,080 different combinations to explore. To reduce the computational burden, 100 hyperparameter combinations were randomly drawn from the overall hyperparameter combination space and fit on the training set with five-fold cross-validation, totaling 500 fits. Then a local exhaustive grid search with five-fold cross-validation was run with finer hyperparameter values in a close range around the previously found best hyperparameters also detailed in Table 2. The final hyperparameters after this two-fold process are presented in Table 2. To prevent overfitting by the RF algorithm the hyperparameter search was restricted in range and subjected to five-fold cross-validation. The deeper a decision tree, the more closely it can mirror the training set as the decision tree becomes increasingly complex and leads to overfitting. As such, the maximum depth of decision trees within the RF was restricted to a maximum of 90 trees in the second search. The number of decision trees in the forest was allowed to grow to large numbers to reduce variance in the resultant predictions. The minimum number of data points in a node before it could split was five to reduce the complexity of the decision tree. The minimum number of data points in a leaf node was kept to two to reduce overly branching and overfitting the training set. Bootstrapping was always kept as the default to reduce overly memorizing the training set since it only uses a randomly drawn set with replacement of the training set datapoints. The bootstrapping set was intentionally kept at a range below 100% of the training set size in the second search to ensure the entire training set could not be drawn each time. These measures lowered the accuracy from the training set but resulted in higher test set accuracies and a smaller gap between the two accuracies, indicating an absence of overfitting.

FINAL MODEL RESULTS

The final RF model has an overall accuracy of 96% on the test set. Additionally, the dataset is imbalanced and therefore the per category accuracy is a more holistic view of model performance. The prediction accuracy per manifestation level was: 99.4% for category 0, 92.2%

for category 1, 94.3% for category 2, 82.4% for category 3, 89.1% for category 4 and 88.9% for category 5. The RF model performs well in detecting no liquefaction manifestation occurrence with slightly lower accuracy for higher category events.

Table 2. Hyperparameters searched over during the two-stage five-fold cross-validation (CV)

Hyper parameter name	First random CV	Second local CV Final	
	5-fold CV	5-fold CV	Optimal values
Number of decision trees in the forest	[200, 400, …,2000]	[200, 400, 600, 800, 1000, 1200]	1200
Max depth of each decision tree	[10, 20, …,80, None (No restriction)]	[60, 70, 80, 90]	90
Min number of data points placed in a node before the node is split	[5, 10]	[5]	5
Min number of data points allowed in a leaf node	[2, 4]	[2]	2
Ratio of train samples to be drawn via bootstrapping	[0.8, 0.9, 1]	[0.8, 0.9]	0.9

Other metrics were also used for model evaluation: True positives (TP), False positives (FP), True negatives (TN), False negatives (FN). As this is a multiclass problem, the one versus rest (OvR) scheme (Pedregosa et al. 2011) is used where a category is termed as “positive” while the rest of the categories are termed as “negative”. The confusion matrix of Figure 4a summarizes the classification accuracy of the final model across each manifestation category. The diagonal cells show the number of TP out of the predicted number of positives along with the accuracy percentage for the correct category predictions. The misclassified prediction cells contain both the number of incorrect predictions and the percentage of the given category in the incorrect prediction category. The confusion matrix shows a slight tendency for underprediction, with more incorrect predictions into lower manifestation categories than higher manifestation categories. The high accuracy for no manifestation is useful for determining whether mitigation for liquefaction is needed, but the slight underprediction for the other categories can underestimate the extent of potential damage. However, the predictions can potentially underpin and propel detailed investigations and predictions depending on the consequence and risk assessment for a structure at a certain location.

Further examination of model accuracies can be found via the receiver operating characteristic (ROC) curves and the precision-recall curves (PRC) in Figure 4c and 4d. The True Positive Rate (TPR) is defined as TP/(TP+FN) and can be interpreted as the amount of correctly labeled observations for a category out of all the observations that were supposed to be in that category. The False Positive Rate (FPR) is defined as FP/(FP+TN) and can be interpreted as the amount of incorrectly labeled datapoints for a category out of all the datapoints that are not in that category. A ROC curve closer to the top-left corner indicates a good fit of the model. The model performs well on each category, with similar ROC curves for all classes indicating that the model is not lacking in prediction of any given category. A ROC curve comes with Area Under the ROC Curve (AUC) values that range from 0 to 1. The greater the AUC value, the better the model is at classifying for the given category. A PRC curve also has an area value that captures a

similar concept. Figure 4c illustrates the AUC values and indicates that the model is skilled at liquefaction manifestation classification with all values at 1. Figure 4d showcases all PRC curves have an area over 0.98, with category 0 at an area of 1 as well.

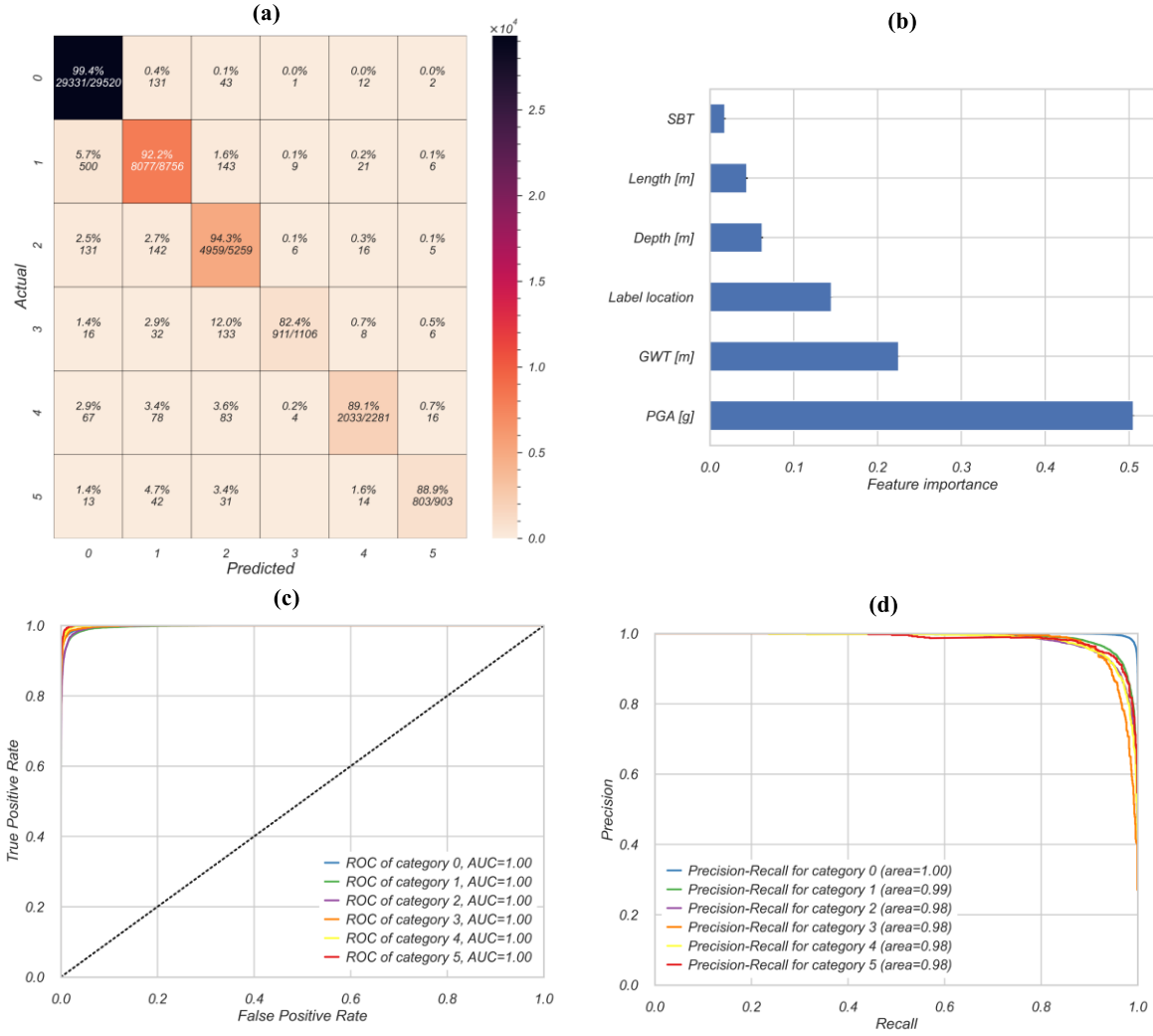

Figure 4. Final RF model metrics on the testing set: (a) Confusion matrix, (b) Feature importance plot, (c) ROC curve, (d) PRC curve. Explanations in surrounding text.

A PRC is useful in the case of an imbalanced dataset as the individual category accuracies can be observed. Precision is defined as TP/(TP+FP) and can be interpreted as how precisely the model will label the correct category for an observation. Recall is defined as TP/(TP+FN) and can be interpreted as how well the model can find all the correct observations for a given category. A PRC curve close to the top-right corner is better. For this model, category 0 is the best performing as to be expected due to the high number of datapoints in this category in the dataset. The rest of the categories perform similarly well without any noticeable drops in precision from one another or along the plot, indicating that the RF model is overcoming the imbalanced dataset and learning patterns from the data instead of purely memorizing the training set.

For interpretability purposes, a feature importance plot was constructed for the RF model. Feature importance or Gini importance is calculated by using the mean and standard deviation of the accumulated Gini impurity decreases for each feature used in each decision tree. Figure 4b showcases the results, with PGA being the most important variable for predicting the liquefaction manifestation. Next in importance is the GWT, which captures the amount of unsaturated material within the CPT location which has an influence on whether the liquefaction could even manifest. Then there is the location label, which indicates that certain clusters of geographical and thus geological/depositional areas are more prone to liquefaction. This was expected due to the prior knowledge that saturated sandy soil is prone to liquefaction, so areas around pockets of water would be more prone to it. The less influential variables are depth and thickness of the soil layer, with depth being slightly more important. As even a small layer of liquefiable soil could cause liquefaction manifestation, the thickness of the soil layer not being very influential reinforces the observations of others. Surprisingly, SBT does not seem to have a major predictive effect on the occurrence of liquefaction manifestation, suggesting that it has an influence, but the PGA, GWT, and location label are more important for a holistic prediction of liquefaction manifestation

CONCLUSIONS

The ability of the Random Forest classification algorithm to predict multiclass liquefaction manifestation was investigated using the Geyin et al. (2021) database of observations from three earthquakes in New Zealand. The algorithm performed successfully yielding 96% overall accuracy in predicting different categories of manifestation on the test dataset. It was found that the key to the RF algorithm's success was robust feature engineering that drew from prior geotechnical earthquake engineering knowledge that worked to preserve the original data's physical constraints as liquefaction manifestation is influenced by both stratigraphy and geology. CPT data is commonly used in empirical methods and ML algorithms but reframing the available data into a format suited specifically for the RF algorithm required consideration of how each feature's representation would influence the algorithm's accuracy and efficiency. The RF algorithm's potential for overfitting was addressed by restricting the hyperparameter values allowed in two stages of grid searches with five-fold cross-validation to find the optimal hyperparameters.

Feature engineering by preserving geological stratification can become a novel approach where only sections of a CPT are used along with corresponding features to obtain localized information on whether liquefaction manifestation will occur. Keeping the soil layers distinct can also allow for isolation of the most influential soil type for liquefaction manifestation or for identifying interaction effects. This breakdown of a CPT into soil layers can also allow for further potential integration with geospatial probabilities per soil type as prior distributions and resulting in a probabilistic instead of deterministic liquefaction manifestation outcome.

The incorporation of geographical locations as Northing-Easting labels currently makes the model only viable for the area the location grid covers. However, this location label feature indirectly captures broad geological indicators such as proximity to bodies of water or known swaths of alluvial soil and can underpin extension of this work by accounting for geology more directly than with a location label. Future work can explore the above suggestions or try new representations of the same dataset's features to explore how changes in feature engineering and algorithm choices influence the predicted results and accuracies.

REFERENCES

Bassal, P. C., Boulanger, R. W., and DeJong, J. T. (2022). System Response of an Interlayered Deposit with Spatially Distributed Ground Deformations in the Chi-Chi Earthquake. *Journal of Geotechnical and Geoenvironmental Engineering*, 148(10), 05022004.

Bengfort, B., et al. (2018). *Yellowbrick (0.9.1)*.

Beyzaei, C. Z., Bray, J. D., van Ballegooy, S., Cubrinovski, M., and Bastin, S. (2018). Depositional environment effects on observed liquefaction performance in silt swamps during the Canterbury earthquake sequence. *Soil Dynamics and Earthquake Engineering*, 107, 303–321.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Cubrinovski, M., and Robinson, K. (2016). Lateral spreading: Evidence and interpretation from the 2010–2011 Christchurch earthquakes. *Soil Dynamics and Earthquake Engineering*, 91, 187–201.

Cubrinovski, M., Bray, J. D., Taylor, M., Giorgini, S., Bradley, B., Wotherspoon, L., and Zupan, J. (2011). Soil Liquefaction Effects in the Central Business District during the February 2011 Christchurch Earthquake. *Seismological Research Letters*, 82(6), 893–904.

Cubrinovski, M., Rhodes, A., Ntritsos, N., and Van Ballegooy, S. (2019). System response of liquefiable deposits. *Soil Dynamics and Earthquake Engineering*, 124, 212–229.

Durante, M. G., and Rathje, E. M. (2021). An exploration of the use of machine learning to predict lateral spreading. *Earthquake Spectra*, 875529302110046.

Geyin, M., Maurer, B. W., Bradley, B. A., Green, R. A., and van Ballegooy, S. (2021). CPT-based liquefaction case histories compiled from three earthquakes in Canterbury, New Zealand. *Earthquake Spectra,* 37(4), 2920–2945.

Geyin, M., Maurer, B. W., and Christofferson, K. (2022). An AI driven, mechanistically grounded geospatial liquefaction model for rapid response and scenario planning. *Soil Dynamics and Earthquake Engineering*, 159, 107348.

Harris, C. R., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362.

Maurer, B. W., Green, R. A., Cubrinovski, M., and Bradley, B. A. (2015). Assessment of CPT-based methods for liquefaction evaluation in a liquefaction potential index framework. *Géotechnique*, 65(5), 328–336.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt and J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61).

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Robertson, P. K. (2010). Soil behaviour type from the CPT: An update. *2nd International Symposium on Cone Penetration Testing*, (pp. 56).

Van Ballegooy, S., Malan, P., Lacrosse, V., Jacka, M. E., Cubrinovski, M., Bray, J. D., O'Rourke, T. D., Crawford, S. A., and Cowan, H. (2014). Assessment of Liquefaction-Induced Land Damage for Residential Christchurch. *Earthquake Spectra*, 30(1), 31–55.