Use of Tree-Based Machine Learning Methods for Stratigraphic Classification in 3D Geological Modelling

Yan Wei¹, Zhu Xing¹,²,*, Chu Jian¹*, Wang Kangda¹, Wu Shifan¹, Kiefer Chiam³

¹. School of Civil and Environment Engineering, Nanyang Technological University, Singapore 618798, Singapore
². State Key Laboratory Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China
³. Building and Construction Authority, Singapore.

*Corresponding: cjchu@ntu.edu.sg; xing.zhu@ntu.edu.sg

Abstract. The need for underground development has been expanded due to rapid urbanization and to reduce the negative impact on city living. In recent years, 3D geological modelling has been widely used by engineers and geo-scientists for desk study and has shown its capability of integrating geological and geotechnical information for better usage in building and civil engineering projects. To provide techniques and instruction, a 3D Geo-data Modelling and Management System (GeM2S) has been established to better understand the subsurface conditions in Singapore. However, the geological stratum between the existing boreholes is often not investigated, which brings the possibility of vital errors in underground design due to inaccurate interpretation of the ground conditions. It is desirable to utilize advanced machine learning methods to predict and update the 3D geo-models. Machine learning is regarded as a subset in the field of artificial intelligence, which has shown its rapid development recently. However, the accuracy of the model is one of the major concerns when the techniques are applied. In this study, four machine learning models are proposed to provide the solutions for stratigraphic classification, namely, Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Categorical Boosting (Catboost). The borehole data are located in the Bishan planning region in Singapore. The voxel-based input data for the models are spatial coordinates (X, Y, Z) and ground surface elevation. The prediction results demonstrate that LightGBM with optimization have produced the highest performance in this multi-classification problems. Finally, based on the prediction results of LightGBM, the voxel-based 3D geological models and the selected cross-sections are built for further analysis and comparison.

1. Introduction
The urbanization of smart cities has led to a rapid development in data integration techniques. With the increase trend of exploring better use of underground space, it is important to conduct adequate site investigation to understand the subsurface conditions, which may incur a large amount of investigation cost. Over the decades, with the help of geological modelling software such as GoCAD [1, 2], stratum between the investigated holes could be generated using traditional interpolation techniques [3, 4] and this has been illustrated in published papers by others. To provide a platform for use by engineers and
planners, a 3D Geo-data Modelling and Management System (GeM2S) is created. This is a web-based 3D design tool for managing boreholes information and use for visualization of the subsurface environment for potential underground projects in Singapore[5]. However, the limitation of the 3D geological model is obvious: the condition of the gap among the existing boreholes is unknown, which may bring risks in underground design and construction.

In this case study, machine learning was used in 3D geological modelling for stratigraphic identification. The use of machine learning may be the solution for reducing the uncertainties in 3D geological modelling. To improve the interpolation precision of formation and their interfaces, researchers[6] have attempted to predict the stratum using RNN with pre-processed borehole information. Alternatively, 3D geological modelling could be regarded as layer classification, which has shown to be workable using support vector machine [7]. A voxel-based model [8] was introduced with a given serial number with spatial coordinates, where individual voxels were inputted into the KNN model and neural network classifier. The results showed that KNN provided a better accuracy with local nature and acceptable voxel distribution.

To further investigate the capability of machine learning, four tree-structure models with Bayesian optimization are proposed, focusing on the stratigraphic classification. Based on the 543 boreholes information gathered from the Bishan region in Singapore, the 3D geological model could be reconstructed using these machine learning methods with a higher precision.

2. Data and Methodology

2.1. Geological data

As shown in Figure 1, Bishan is situated near the center region of Singapore. The 543 boreholes are distributed across the study area, with maximum depth up to 65 meters below ground level. In this study, the geological formations are represented by codes that indicate a different property for each code. There are eight different formation codes used (Figure 1): Top layer (FILL), Kt (Transitional Member), Ka (Alluvial Member), Km (Marine Member), OA (Old Alluvium), J (Jurong Formation), and BT (Bukit Timah Granite), and R (Rock). Kt, Ka, and Km are the layers described as sedimental mud or sand. OA is described as loose to dense coarse quartz-feldspar sand and gravel, while J is of sedimentary rocks origin. BT is defined as granite containing weathering grade from IV above, while code R is
introduced to represent the granite voxels holding a weathering grade from I to III. The original borehole information is converted into voxels, which are defined as points in the space as it progresses in depth, with its three-dimensional coordinate (x, y, z), ground surface elevation, and its specific property (geological formation code). The coordinate and ground surface are selected as inputs to estimate the stratum property as outputs. After normalization of the inputting data, it is essential to note that the data are split into 80% for training and 20% testing sets in terms of borehole units rather than voxels.

2.2. Random Forest (RF)

Random Forest algorithm was proposed by Breiman [9], which has obtained wide attention among researchers in 2010 and surpassed the utility of the kernel approach. RF is the ensemble learning of tree predictors. In other words, RF is a classifier with the collection of tree-structure classifiers \{h(x, \theta_k), \ k = 1, 2, \ldots \}, where \theta_k is the independent distributed random vector [9], and integrating into a single prediction by taking the average. To put it more intuitively, RF can be described as:

\[
\text{Random Forest} = \text{Decision Tree} + \text{Bagging}
\]

2.3. Extreme Gradient Boosting (XGBoost)

To further develop the performance of gradient tree boosting, XGBoost, proposed by Chen and Guestrin [10], is a scalable tree learning system based on the improvement of (GBDT). XGBoost is also the learning ensemble of a few weak learners into a strong learner with boosting process, which takes the advantages of regularization, parallel processing, and Classification And Regression Tree (CARTs), to prevent overfitting [11]. In other words, XGBoost is an optimization of GBDT for engineering used.

As other ensembled learners, the output of the model \(\hat{y}_i\) is the collection of various decision trees by voting or average, namely additive function, which can be illustrated as:

\[
\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F
\]

Where \(f_k(x_i)\) is the result of the \(k^{th}\) tree with the training set \(x_i\).

2.4. Light Gradient Boosting Machine (LightGBM)

As a neoteric method proposed by Ke, Meng [12], Light Gradient Boosting Machine has rapidly become one of the most powerful members in Gradient Boosting Decision Tree (GBDT). It has been observed that in the real case, when machine learning applied in industry, high dimension and large data have encountered troubles in the low-efficiency training process. Hence, LightGBM has an absolute advantage in fast training with higher accuracy, which proved to be useful in many disciplines.

With the aim of reducing time in traversal times, a histogram-based algorithm is replaced by the data structure in XGBoost as bin values for every feature. Each histogram gains information of gradient sum and the number of bins, hence the maximum gain can be obtained by comparing the difference of bins in father node and child nodes as function. This process will reduce the memory to only an eighth of its original value.

2.5. Categorical Boosting (Catboost)

Developed by Prokhorenkova [13] from Yandex, Russia. Categorical Boosting (Catboost) has been proposed as a state-of-the-art boosting toolkit which has proven with high accuracy among various
datasets. The major differences between Catboost and other boosting machines are in three aspects: oblivious tree-based structure, ordered target statistics for categorical features, and ordered boosting approach. Unlike the traditional decision tree for LightGBM and XGBoost, the oblivious tree conducts the same splitting criterion at the same level. Thus, this structure can better solve the overfitting problem because of the balance when many trees are trained [13]. Moreover, it will speed up the process of prediction by assigning index values (1 or 0) for each level. In this process, going through the whole tree only requires the reading of the index array rather than complex criteria to conduct the judgment.

3. Results and Discussion

3.1. Performance of Machine Learning Models

| Model          | Manual-tuning          | Bayesian opt.          |
|----------------|------------------------|------------------------|
| RF             | Prec. | Rec.  | F1-score | Prec. | Rec.  | F1-score |
| Manual-tuning  | macro avg | 0.695 | 0.676 | 0.677 | weighted avg | 0.766 | 0.780 | 0.768 |
|                | weight avg | 0.731 | 0.675 | 0.685 | weighted avg | 0.779 | 0.789 | 0.775 |
| Bayesian opt.  | macro avg | 0.703 | 0.678 | 0.680 | weighted avg | 0.768 | 0.778 | 0.767 |

| Model          | Manual-tuning          | Bayesian opt.          |
|----------------|------------------------|------------------------|
| Manual-tuning  | Prec. | Rec.  | F1-score | Prec. | Rec.  | F1-score |
| LightGBM       | macro avg | 0.684 | 0.678 | 0.666 | weighted avg | 0.775 | 0.781 | 0.772 |
|                | weight avg | 0.708 | 0.680 | 0.686 | weighted avg | 0.774 | 0.783 | 0.774 |
| Bayesian opt.  | macro avg | 0.693 | 0.656 | 0.658 | weighted avg | 0.765 | 0.774 | 0.763 |
|                | weight avg | 0.706 | 0.691 | 0.683 | weighted avg | 0.773 | 0.781 | 0.771 |

The evaluation report is presented in Table 1. According to the table, the Bayesian-tuning models present 1% to 3% better in the testing set in comparison to the manual-tuning models. Among all the improvements, XGBoost has less than 1% increase in F1-score while Catboost and LightGBM has a raise score of 2% to 3%.

Among the compared models, the F1-score of LightGBM demonstrates an outstanding performance, indicating a macro average of 0.686 and a weighted average of 0.774. RF comes close to LightGBM with a macro average of 0.685 and a weighted average of 0.775. Furthermore, the precision of RF is even slightly higher than LightGBM. Catboost comes close to the third that showing an F1-score of 0.683 for macro average and 0.771 for weighted average. However, the macro average recall of Catboost has outperformed others, indicating high accuracy in the true value for minority groups. XGBoost has poorer performance as it has a score of 0.68 and 0.767 for macro average and weighted average, respectively. Figure 2 illustrates a visualization of prediction results for each formation, where some of the Kt formations are wrongly predicted as BT, and the R formation is easy to be predicted as BT as
well. It appears that the predictions will be rather challenged when the positions of Kt and R are very close to the BT. J formation in this area is situated in the southwest corner replacing the formation of BT, thus the simple location of J makes high recall in all predictions. The overview of the confusion matrix shows that RF is somewhat better than others in terms of BT (0.91), OA (0.93), and Km (0.62), while Catboost is higher in R (0.67), and FILL (0.7). LightGBM is good at predicting Kt, even if it is only 0.3 of recall. Regarding the comparison shown in Figure 3, the precision of OA, BT, and R, are stably achieved in all models, while Kt is rather hard to estimate. The overall precision shows that RF is better than others in terms of OA (0.834), Kt (0.55), Ka (0.733), and Km (0.837), while Catboost is higher in FILL (0.64). LightGBM is good at predicting J (0.746), BT (0.82), R (0.788).

![Figure 2. Performance of models for each formation](image)

Figure 2. Performance of models for each formation
3.2. Voxel-based 3D Geological modelling

In accordance with the time for training and the quantified analysis above, the efficiency of LightGBM in practice outperformed other methods with relatively high evaluation results. It revealed that LightGBM can be chosen as the reference model for real-world cases. Thus, in this study, the 3D geological model in Bishan can be re-constructed in LightGBM based on the modelling solutions. The voxels observed and predicted by LightGBM are scattered in a 3D plot, as shown in Figure 4(a) and Figure 4(b). The main geological formations in Bishan are BT (cyan) and OA (red). The former can be observed at the northwest part of Bishan while the latter is mostly appeared at the south-eastern part. The top is covered by FILL and other formations. The distribution of J formation (green) is mainly located at the southern corner of the area. Under the specific perspective of this view, the predicted model is very similar to the observed one. The difference of the voxels can be distinguished as marked with boxes, most of which are wrong predictions of FILL and Ka formation.

To better illustrate the constructed model in prediction, a selected cross-section is extracted as 2D view for visualization in Figure 5(a) for observed model and Figure 5(b) for predicted model. It can be concluded from the figures that the prediction of the stratum reveals the distribution of formations along with the depth. For example, the Kt formation in dark red is found to be overlying the Ka in yellow and Km in light blue which both are intercepting with each other. The model is able to capture the geometric trend of the geological formation in Bishan case. However, there are obvious variations in the geological formation of Kt, Km, Ka, and OA, as marked with boxes. It can be observed that the shape of geological formation in the predicted model is relatively uniform and regular that fails to capture the complex geometry of intersection forms in practice. The results may be affected by the distribution of formation as well as the quantity of voxels at the other cross-sections.
Figure 4. The (a) observed and (b) LightGBM-predicted 3D geological models for Bishan
4. Conclusion

This study investigated the use of four tree-structure machine learning methods, namely Random Forest, XGBoost, LightGBM, and Catboost for stratigraphic classification of 3D geological model in Singapore. The existing borehole data were interpreted as voxels, while the separation of training and testing set was based on the borehole units. The study demonstrated that the voxel-based data in machine learning can be used for the prediction of the stratum at the unmeasured area with the input data - coordinates (X, Y, Z) and ground surface. Overall, the results demonstrated that LightGBM was of high performance with 0.686 for macro average and 0.774 for weighted average after hydra-parameters tuning, which seems to be a good choice for application in the real-world cases. The voxel-based 3D geological models predicted by LightGBM were built to reveal the distribution of the formation that could reflect the soil distribution at the existing location where boreholes are located. The findings also suggest a relatively low score in minority classes (such as Kt formation) due to the insufficient of borehole data which is a common challenge in real practice that needs to be tackled during the data collecting and cleaning process. The performance of the machine learning models could be further improved with more evenly data distribution.
5. Acknowledgement

This research is supported by National Research Foundation (NRF) Singapore, under its Virtual Singapore Programme (Grant No. NRF2019VSG-GMS-001) and Ministry of National Development (MND) Singapore, under the Land and Livability National Innovation Challenge (L2NIC) program (Grant No. L2NICFP2-2015-1). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore Ministry of National Development and National Research Foundation, Prime Minister’s Office, Singapore.

The authors would also like to thank Drs Xiahua PAN and Xiahui QI for their contribution to the early part of this research.
References

[1] Kaufmann, O. and T. Martin, Reprint of “3D geological modelling from boreholes, cross-sections and geological maps, application over former natural gas storages in coal mines” [Comput. Geosci. 34 (2008) 278–290]. Computers & Geosciences, 2009. 35(1): p. 70-82.

[2] Jin, X., et al., 3D geological modelling and uncertainty analysis for 3D targeting in Shanggong gold deposit (China). Journal of Geochemical Exploration, 2020. 210.

[3] Marinoni, O., Improving geological models using a combined ordinary–indicator kriging approach. Engineering Geology, 2003. 69(1-2): p. 37-45.

[4] Sun, Y. and G. Wang, 3D Geological Modeling of Pulaing Copper Deposit, Yunnan Province of China, in 2012 International Conference on Communication Systems and Network Technologies. 2012. p. 294-298.

[5] Pan, X., et al., 3D Geological Modelling: A Case Study for Singapore, in Information Technology in Geo-Engineering. 2020. p. 161-167.

[6] Zhou, C., et al., A Stratigraphic Prediction Method Based on Machine Learning. Applied Sciences, 2019. 9(17).

[7] Smirnoff, A., E. Boisvert, and S.J. Paradis, Support vector machine for 3D modelling from sparse geological information of various origins. Computers & Geosciences, 2008. 34(2): p. 127-143.

[8] Jankowski, S., et al., Modeling Engineering-Geological Layers by k-nn and Neural Networks, in Communications in Computer and Information Science. 2017. p. Springer, Cham.

[9] Breiman, L., Random Forests. Machine Learning, 2001. 45: p. 5–32.

[10] Chen, T. and C. Guestrin, XGBoost A scalable tree boosting system, in KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[11] Thongsuwan, S., et al., ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost. Nuclear Engineering and Technology, 2020.

[12] Ke, G., et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017.

[13] Prokhorenkova, L., et al., CatBoost: unbiased boosting with categorical features. 2017.