Recent Trends in Civil Engineering and Built Environment Vol. 2 No. 1 (2021) 551-560

RTCEBE

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/rtcebe
e-ISSN : 2773-5184

Settlement Prediction Model in Consideration of Static Loading on Soft Clay by Utilising Machine Learning Method

Elilragi Ganasan[1], Alvin John Lim Meng Siang[1]

[1]Faculty of Civil Engineering and Built Environment,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/rtcebe.2021.02.01.060
Received 30 January 2021; Accepted 28 April 2021; Available online 30 June 2021

Abstract: The past researchers had deduced that the current condition of marine clay soil in Sabak Bernam, Selangor were highly sensitive soil that is always associated with high settlement and high instability, poor soil properties that are not suitable for engineering requirements. Hence, there is a need to develop an artificial neural network (ANN) model to predict the soil settlement by associating with the current history of soil settlement thus producing a reliable prediction model in the future. The aim of this research is to obtain the best soil settlement ANN model which is selected among these four types of soil settlement prediction models (two from deep learning (DL) and two from support vector machine (SVM)) using konstanz information miner (KNIME) machine learning. SVM - dot kernel prediction model exhibited 34% less discrepancy values between measured and predicted Sabak Bernam, Selangor marine clay soil settlement compared to DL - max-out (58%), DL - rectifier (56%) and SVM - neural (59%) and was therefore chosen. Further optimization was made on SVM-dot model in order to reduce the error between measured and predicted value using data splits and performance indexes. At the end of the analysis SVM-dot with root mean square error (RMSE) performance index has achieved further refinement up to 5% between measured and predicted value hence, this model has been chosen and would be suitable in predicting the settlement of problematic soil.

Keywords: Artificial Neural Network, Deep Learning, Support Vector Machine, Konstanz Information Miner, Root Mean Square Error.

1. Introduction

In the field of soft soil engineering, settlement is defined as the vertical movement of the ground, generally caused by changes in stress within the earth [1]. Settlement of the ground consists of three components. There are immediate settlement (usually called as elastic settlement, although this is a misnomer), consolidation settlement (known as primary settlement) and creep settlement (known as secondary settlement) [2]. Subsidence is a concept that is sometimes used to describe 'caving in' or sinking of the ground which may not be consistent with increases in soil stresses. Excessive settlement

*Corresponding author: alvin@uthm.edu.my

or subsidence can lead to distortion and damage to buildings, facilities and utilities that are dependent on material subject to movement. Settlement may be almost immediate, or it can take years or decades to occur, depending on the underlying soil conditions and the cause of movement [3]. Soft soil areas are being widely used for the development of infrastructure and other associated developments due to the restricted shortages of 'suitable' ground for the construction of infrastructure.

Marine clay is a type of soil that primarily occurs in coastal corridors, lowland and offshore zones, as well as in other sections of the world. It can be defined as soft sensitive soil that is often associated with high settlement and high volatility, weak soil properties that are not appropriate for engineering requirements, inconsistency of results, low unconfined compressive strength between 25 and 50 kPa, and flat or featureless surface [4]. In this research, the site soil area was acquired in Sabak Bernam, Selangor, where the area is roughly 2 km from the coastal plain of Malacca Straits, as in the Generalised Soil Map of Peninsular Malaysia [5]. The authors' analysis showed that Selangor marine clay soils have high pH, high cation exchange capacity (CEC), specific surface area (SSA) values and high clay content with the existence of montmorillonite [6]. Therefore, challenges arise towards engineer facing in all sorts of problem to design and construct foundation of building, road and highway embankment. They are subjected to massive primary and long-term consolidation settlement even when subjected to a moderate load.

The most critical geotechnical challenges are excessive settlement and differential settlement leading to hazardous and discomfort in road usage. There is a need to predict the soil settlement in future by developing the artificial neural network (ANN) model based on the history of soil settlement. This model would be easy for hands-on without much complexity in understanding the rate of settlement change using feasible machine learning. In addition, there is no need much time required for ANN to produce the prediction result compared to the physical testing (required longer period and higher cost consuming). Hence in this research, konstanz information miner (KNIME) software will be used to predict the soil settlement in future by developing the ANN model.

2. Materials and Methods

The total number of 100 data of marine clay soil settlement with related mechanical properties will be assess from the research published report on problematic Sabak Bernam, Selangor marine clay. Then, certain mechanical properties data (listed Table 2) will be selected as to suffice with the settlement rate of Selangor marine clay, hence these data will be the key in developing artificial neural network (ANN) settlement prediction model using konstanz information miner (KNIME) software. In this research, four different types of ANNs prediction models; deep learning (DL) max-out, deep learning (DL) rectifier, support vector machine (SVM) dot and support vector machine (SVM) neural from KNIME software will be opted as a platform that provides an integrated environment for machine learning, data mining, text mining and fast with an easy-to-use visual environment for predictive analytics.

In this work, the proposed framework model in Figure 1 (a) Deep learning and (b) Support vector machine will be applied for regression of settlement rate prediction models. Four types of settlement rate prediction models are proposed as in Table 1 with potential input parameters. The regression performance of classification for all models were evaluated by the labelling data set, where the inputs data (listed Table 2 and as attached in Appendix) were attributed with label role while an attribute with prediction role for existing Selangor marine clay soil settlement data. The following KNIME workflow steps are applied in all four models (DL and SVM) as shown below.

Step 1 : **Download data and *"Create New Workflow"***
First of all, to get started, download and create the settlements data into CSV-file in the form of Excel-file that contains all the data which is going to use in the

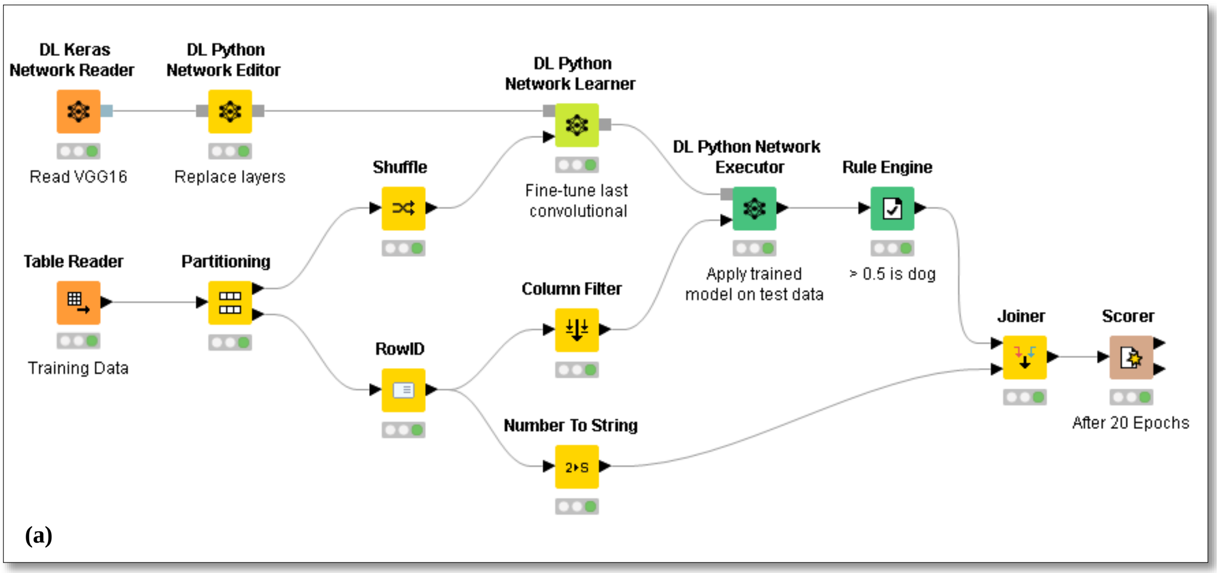
workflow. Open KNIME analytics platform and create a new, empty workflow by clicking *"New"* in the toolbar.

Step 2 : **Drag and drop CSV-file into *"Workbench Editor"***
From the download folder, drag and drop the CSV-file into the *"Workbench editor"*. A file reader node will appear on the workflow editor and its configuration dialog will pop-up.

Step 3 : **Filter data with the *"Column Filter Node"***
To filter some of the columns out, need to use the *"Column Filter node"*. In the node repository panel on the left, write *"column filter node"* in the search field drag and drop the *"Column Filter node"* to the workflow editor and then, connect it in the input port with the output port of the *"File Reader node"*. Then, move of the *"columns"*, of the eight parameters into the *"Include field"* on the right side of the dialog, then click *"OK"*. After executing the node, the filtered data table is available at the output port of the *"Column Filter node"*.

Step 4 : **Execute and open output visualization**
Then, select the model type for example SVM by arrange and connecting them from node 3 to node 4 and to node 5. Execute and view the output of these last two nodes. Right-click the node and choose *"Execute and Open Views"* from the context menu. A new window will open showing the charts/tables which is built with the settlements data.

The role of selecting input variables is common to the creation of most models of regression and depends on the discovery of relationships within the available data to identify acceptable model output predictors. In the case of ANN, however, there is no such inference about the model's structure. The input variables are instead chosen from the available data and the model is subsequently developed. The smaller number in sample size with the less number of inputs are effectively cover the prediction observed in the broader domain where a small variance developed in the regression model based on the smaller number of inputs dataset [7].



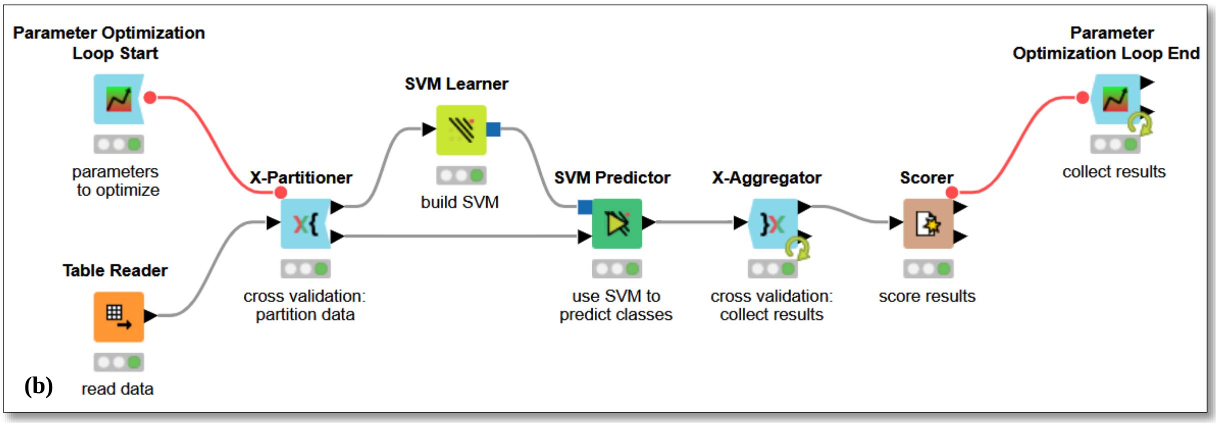


Figure 1: Proposed model framework for (a) Deep learning and (b) Support vector machine

Table 1: Detailed on four types of prediction models

Num	Model	Prediction method
1.	Deep learning (DL)	**Max-out**: Based on the maximum coordinate of the input vector.
2.	Deep learning (DL)	**Rectifier**: Rectifier Linear Unit (RLU) which choose the maximum of (0, x) where x is the input value.
3.	Support Vector Machine (SVM)	**Dot**: The dot kernel is defined by k(x,y)=x*y i.e. it is inner product of x and y.
4.	Support Vector Machine (SVM)	**Neural**: The neural kernel is defined by a two layered neural net tanh (a x*y+b) where a is alpha and b is the intercept constant. These parameters can be adjusted using the kernel *a* and kernel *b* parameters. A common value for alpha is 1/N, where N is the data dimension.

Table 2: List of data opted as potential input parameters for ANN modelling purpose

Num	Parameters obtained based on previous studies	Abbrev.	Unit
1.	Loading (including variables)	w	kN
2.	Moisture content	M_c	%
3.	Plasticity index	PI	-
4.	Specific gravity	G_s	-
5.	pH value	pH	-
6.	Compression index	C_c	-
7.	Coefficient of consolidation	C_v	mm^2/min
8.	Settlement rate (prediction value)	Δ	mm/min

The label attribute stores the actual observed values whereas the prediction attribute stores the values of label predicted by the regression models. Then, these trained models were provided with testing set for predicting the approachable accuracy of settlement rate data. In order to carried out the simulation, eight numbers of potential inputs vector will be selected with various split data ratios (partitioning for DL and cross partitioning for SVM) of 70:30, 75:25 and 80:20 (training : testing), thus will be applied for all prediction models. At the end of this research, this marine clay soil settlement prediction based on Sabak Bernam, Selangor site would be beneficial as possibly to avoid high costing due to implementation of physical experimentation in the future.

3. Results and Discussion

Four types of soil settlement prediction models were proposed and stated as in Table 1 with potential input parameters as stated in Table 2 and the total of 100 dataset as attached in Appendix. The plots of regression analysis for distribution of predicted soil settlement in all prediction models are provided in Figure 2.

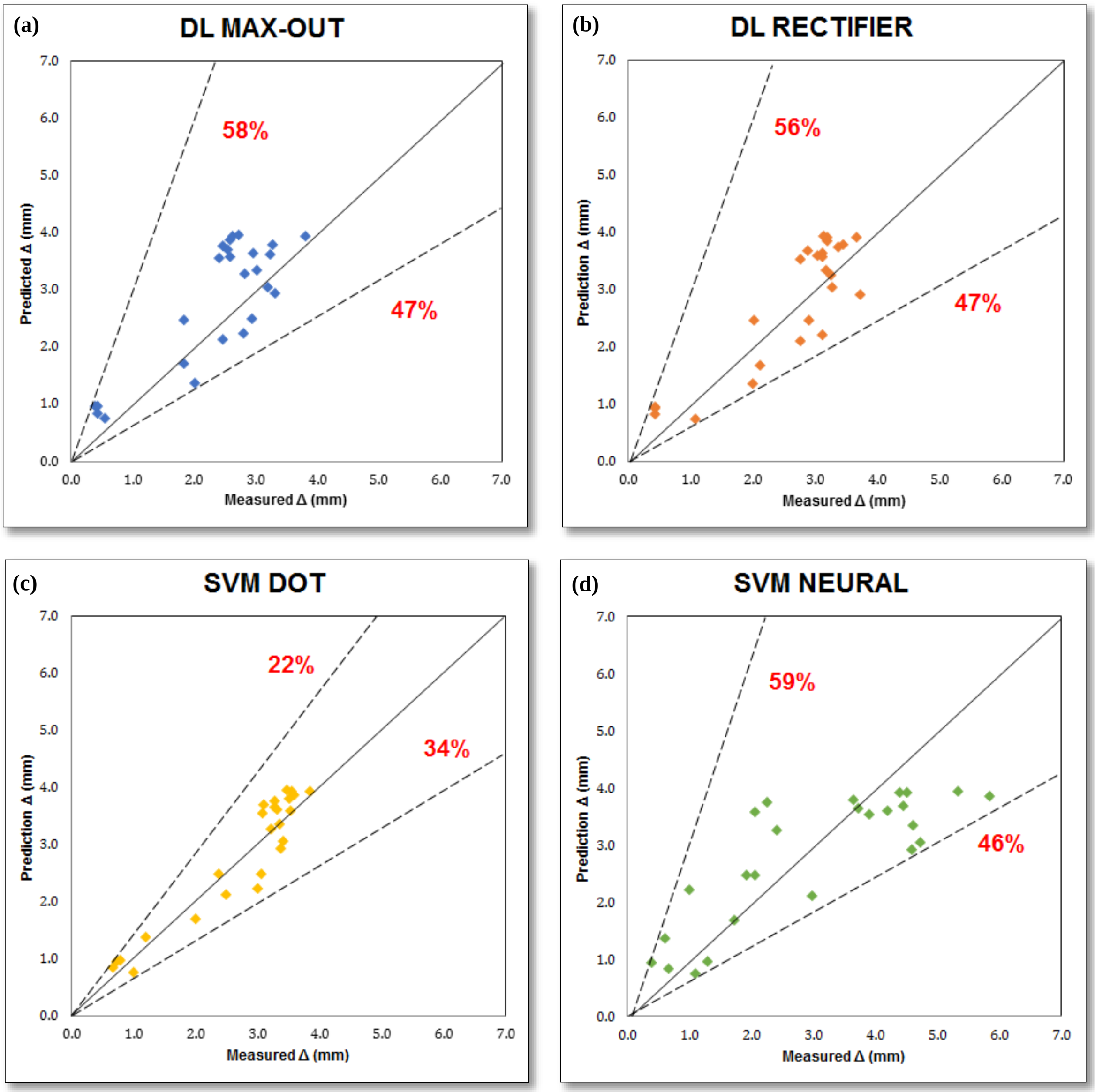


Figure 2: The distribution of predicted marine clay settlement for (a) DL max-out prediction model (b) DL rectifier prediction model (c) SVM dot prediction model and (d) SVM neural prediction model

The Figure 2 includes equity line as a guide, which for the predicted and measured Sabak Bernam, Selangor marine clay soil settlement reflects the state of equal value. The analysis shows that SVM model with dot prediction method presented better prediction with lowest error in which, the distribution of predicted was below (34% error) and above (22% error) of the reference line compared to DL max-out (47% and 58%), DL rectifier (47% and 56%) and SVM neural (46% and 59%). DL max-out and DL rectifier showed that almost all predicted points were distributed around the measured Sabak Bernam, Selangor marine clay soil settlement with highest errors that led to the over-fitting. These high errors implied that insufficient amount of data may had been the main factor as these DL models required a large number of data for complex supported calculation/simulations as supported in past studies [8-9].

The SVM - dot model was selected for further analysis of optimization using data splitting and performance index and is described in the following sub-section. When a specific test set was not available, the split validation operator was implemented to predict the model's fit to a hypothetical test set. The split validation operator can also train on one set of data and check on another set of specific test data. The purpose of splitting data into two different categories in this prediction model was to avoid over- and under-fitting and only optimizing the training dataset accuracy. Hence, there a need of a model that performs well on dataset that it has never seen (test data), which is called as generalization. Figure 3 shows three different results based on linear sampling for split ratio between training and testing of 70:30, 75:25 and 80:20 from the prediction model.

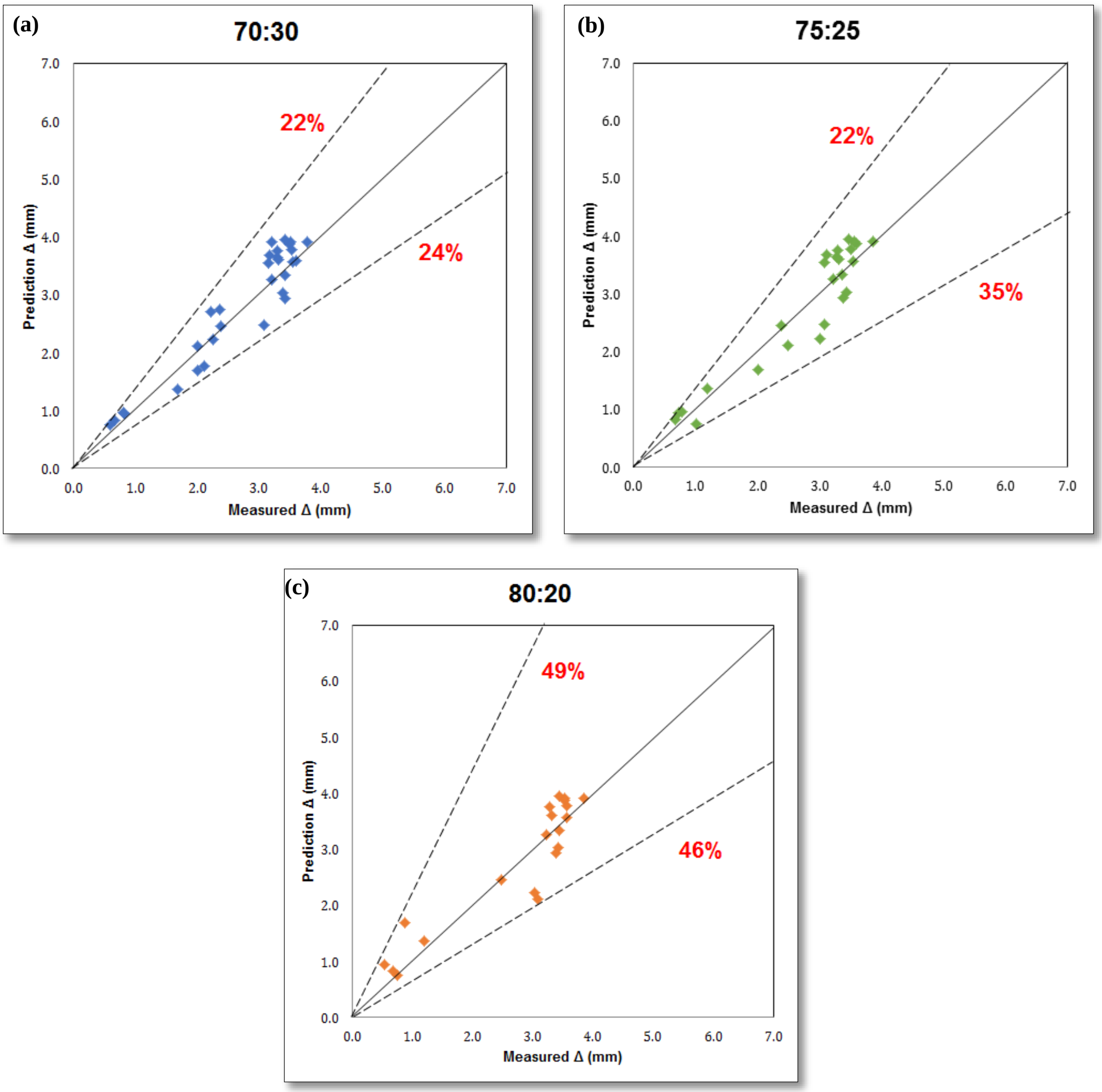


Figure 3: The distribution of predicted results according to every split ratio (a) 70:30, (b) 75:25 and (c) 80:20

Results showed the split ratio of 70:30 with lowest error of predicted Sabak Bernam, Selangor marine clay soil settlement (24%) near the reference line. It is thus a better split ratio for the SVM - dot than 75:25 (35%) and 80:20 (49%) of split ratios. Most of the plotted predicted points for 80:20 split ratio were far below equity/measured line due to unbalanced data from smaller number of dataset thus, providing high variance in prediction, which can significantly change testing accuracy. In other words, significant under-fitting in the 80:20 split ratio may cause redundancy in experimental output data. Past studies have shown that the proportion chosen in the analysis for fewer numbers of dataset (≤ 100 data) was 70% for the training set and 30% for the test. The idea is that more training data is preferred because it makes classification and regression model better whilst more test data makes the error estimate less accurate [10-12]. For this dataset, the 70:30 split is within this experimental range and is a reasonable choice. The trade-off is simple as less the testing data, bigger the variance performance of model algorithm, while more the training data, smaller would be the variance in parameter estimates. From this 70:30 split ratio graph, further optimizations are analysed and discussed based on type of performance indexes (regression).

Performance index operators were used to test the regression task with statistical performance and provided a list of the regression task's accuracy criteria values. The operator output (regression) was selected as it decides the type of learning task and measures the most common requirements for that category automatically. Regression operation is also a method used for numerical analysis and is a statistical measure that evaluate the intensity of the relationship between a dependent variable (label attribute) and a set of other changing variables known as independent variables (regular attributes). In order to evaluate the statistical efficiency of the regression model, the data set must be labelled and must have a label function attribute and a predictive role attribute. The attribute of the label stores the actual observed values and the attribute of the prediction stores the label values predicted by the regression model under discussion. For this study, three types of index regression as in Table 3 were used to refine and reduce the difference/gap between measured and predicted Sabak Bernam, Selangor marine clay soil settlement, and are shown in Figure 4.

Table 3: Types of performance index (regression) used for SVM - dot model

Num	Parameter	Description
1.	Absolute error (AE)	AE is determined by adding the variance of all expected values from the label attribute's actual values and dividing the amount by the total number of predictions.
2.	Prediction average (PA)	PA is determined by adding all the real label values and dividing the total number of examples by this figure.
3.	Root mean square error (RMSE)	RMSE is a quadratic scoring method calculating the error's average magnitude. It is the cumulative square root of variations between predictive and real observation.

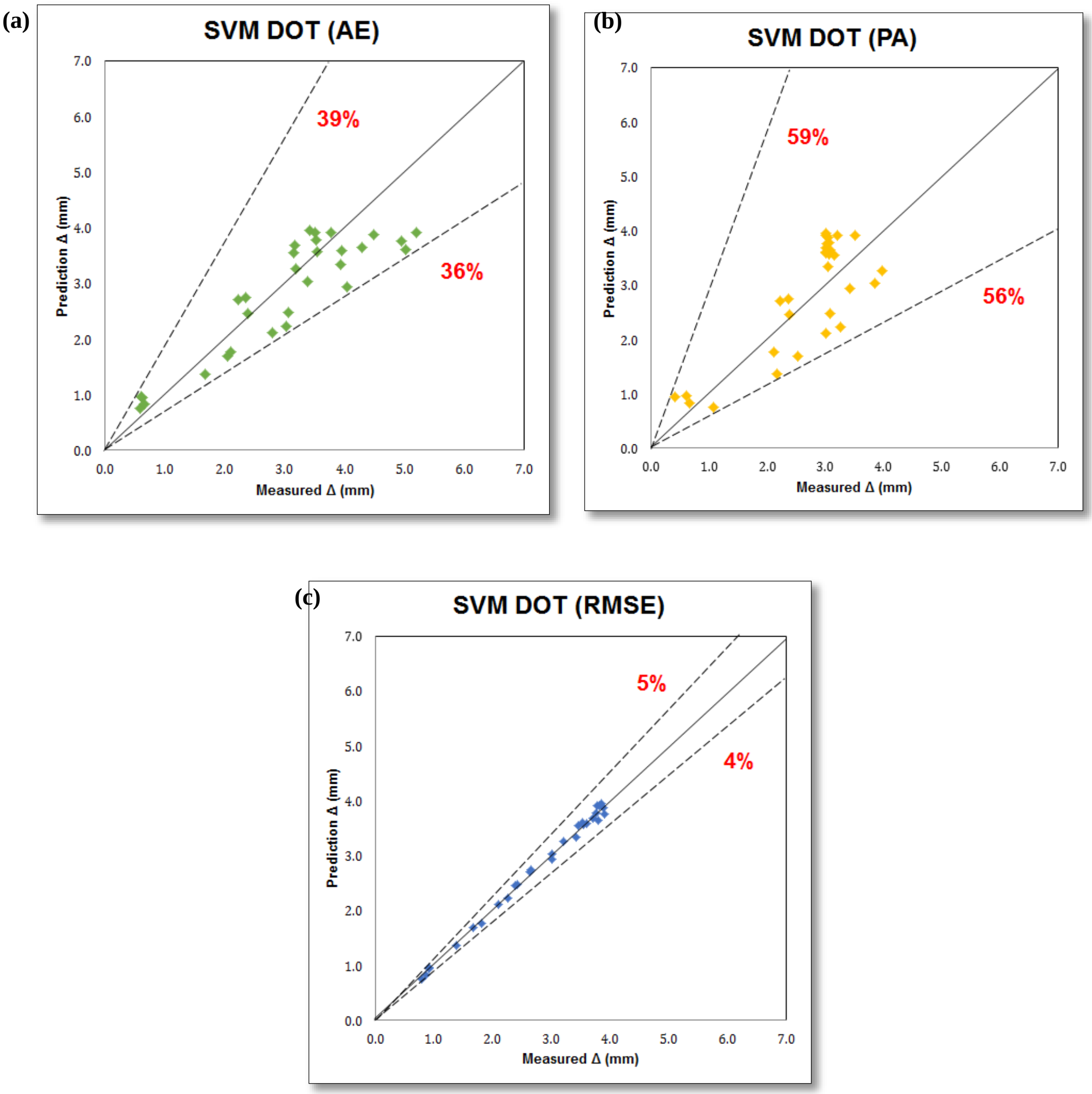


Figure 4: The distribution of predicted soil settlement for (a) Absolute error (AE) (b) Prediction average (PA) and (c) Root mean square error (RSME)

From the Figure 4, RMSE can be seen to exhibit a more consistent standard deviation of residuals (prediction errors) up to (4% - 5%) approaching actual soil settlement compared to AE (36% - 39%) and PA (56% - 59%). The inconsistent prediction performance found over-fitting in the case of AE and under-fitting in the case of PA from the reference lines, respectively. The extent of error found in the prediction model associated with AE and PA may be considered from the point of view of accuracy. The accuracy of a measurement reflects the error or variance of the measurement from the average of a large number of measurements of the same quantity, whereas the precision of a measured value expresses the deviation of the measurement from the real quantity value. Error is viewed from the point of view of precision when the true value is known, but it must be used instead of accuracy when the true value of a quantity is not known.

4. Conclusion

It is concluded that, SVM-dot with adding RMSE performance index using KNIME machine learning shows a significant accuracy prediction of marine clay soil settlement as in experimental result.

Throughout this research, the prediction of the ANN model is favorable as the less differences which was within 5% was found between the prediction and the actual results for the Sabak Bernam, Selangor marine clay soil settlement. Hence, this method would be reliable, easy hands-on and can be continuously improved and develop for other types of problematic soil. Other than that, not much expenses or time needed for running the simulation compared to experimental work which is not very efficient in both cost and time wise.

Appendix

Selected inputs data of marine clay obtained from the settlement reports (2008-2018) in Sabak Bernam, Selangor

Num	Input Parameters							Measured Data
	w (kN)	M_c (%)	PI	G_s	pH	C_c	C_v (mm^2/min)	Δ (mm)
1	70	72	37.2	2.21	4.8	0.623	1.76	0.072
2	68	75	42.3	2.21	4.8	0.651	1.7	0.097
3	68	77.7	41.1	2.21	5	0.63	1.71	0.121
4	90	80	40.9	2.21	5	0.52	1.8	0.325
5	95	76	47.4	2.21	5	0.625	1.83	0.346
6	67	69.9	41.8	2.21	5	0.614	1.69	0.478
7	100	75.6	42.3	2.21	5.2	0.59	1.86	0.742
8	72	87.6	47.3	2.21	5.2	0.83	1.75	0.749
9	75	76.3	40.7	2.21	4.8	0.69	1.77	0.645
10	95	75.5	43.7	2.21	4.8	0.633	1.84	0.83
11	72	76	46.6	2.21	4.8	0.585	1.72	0.947
12	69	79.1	42.9	2.21	4.8	0.601	1.68	0.958
13	90	76.3	38.3	2.21	4.8	0.637	1.8	0.896
14	90	84.5	46.2	2.21	5.2	0.635	1.8	1.76
15	90	77.2	39.4	2.21	5.2	0.64	1.8	1.33
16	87	69.7	43.8	2.32	5.2	0.631	1.79	1.73
17	75	71.2	51.9	2.32	5.2	0.581	1.74	1.07
18	71	70.3	43.1	2.32	5.2	0.574	1.71	1.706
19	73	77.2	45.9	2.32	5	0.601	1.73	1.693
20	100	75.4	48	2.32	5	0.581	1.86	1.352
21	90	71.8	48	2.32	5	0.609	1.8	1.667
22	90	68.5	46.1	2.32	5	0.644	1.8	1.361
23	90	69.9	49.7	2.32	5	0.625	1.8	2.223
24	90	71.1	47.8	2.32	5.2	0.601	1.8	2.133
25	84	73.8	44.2	2.32	5.2	0.636	1.77	2.177
26	76	74.1	43.6	2.32	5	0.624	1.72	2.467
27	70	76.5	40.7	2.32	5	0.575	1.69	1.971
28	88	69.8	39.8	2.32	5	0.612	1.78	1.903
29	86	70.3	45.7	2.32	5.2	0.617	1.77	2.698
30	100	75.4	46.5	2.32	5.2	0.613	1.85	2.199
31	81	72.1	44.8	2.32	5.2	0.6	1.74	2.51
32	76	70.7	41.2	2.32	4.8	0.605	1.7	1.769
33	73	74.3	44.1	2.32	4.8	0.624	1.72	2.741
34	90	75.4	44.1	2.35	4.8	0.605	1.8	2.477
35	90	80.3	48.2	2.35	4.8	0.576	1.8	3.336
36	90	79.3	47.1	2.35	4.8	0.635	1.8	3.169
37	90	74.6	44.3	2.35	4.8	0.584	1.8	2.991
38	90	77.9	43.8	2.35	5.2	0.596	1.8	2.225
39	90	80.1	43.7	2.35	5.2	0.634	1.8	3.154
40	88	80.1	45.4	2.35	5	0.612	1.79	3.042
41	85	80.2	47.2	2.35	5	0.612	1.77	3.411
42	79	75.4	40.5	2.35	5	0.604	1.74	2.117
43	79	79.5	43.4	2.35	5	0.573	1.75	3.018
44	76	78.4	45.2	2.35	5.2	0.58	1.73	3.317
45	90	73.2	41.1	2.35	5.2	0.593	1.8	2.879
46	90	76.1	43.9	2.35	5.2	0.61	1.8	2.778
47	90	80.2	48.4	2.35	5.2	0.614	1.8	3.019
48	90	80.4	47.3	2.35	5	0.611	1.8	3.114
48	90	80.2	45.7	2.35	5	0.619	1.8	3.152
50	90	79.4	45.6	2.35	5	0.617	1.8	3.174
51	75	74.3	47.1	2.35	5.2	0.623	1.77	3.241
52	95	75.4	44.3	2.35	5.2	0.651	1.77	3.262
53	72	80.3	43.8	2.35	5.2	0.63	1.77	3.296
54	69	79.3	43.7	2.35	5.2	0.52	1.75	3.321
55	90	74.6	45.4	2.35	5.2	0.625	1.8	3.389
56	90	77.9	47.2	2.35	5.2	0.614	1.8	3.411
57	90	80	40.5	2.35	4.8	0.59	1.8	3.475
58	87	76	45.9	2.35	4.8	0.83	1.75	2.931
59	75	69.9	48	2.35	4.8	0.69	1.72	3.317
60	71	75.6	48	2.35	4.8	0.633	1.72	3.345
61	73	87.6	46.1	2.35	4.8	0.585	1.72	3.388
62	90	76.3	49.7	2.35	4.8	0.601	1.72	3.462
63	90	75.5	47.8	2.35	4.8	0.637	1.72	3.517
64	90	76	38.3	2.35	5	0.635	1.72	3.547
65	90	79.1	46.2	2.35	5	0.601	1.72	3.642
66	90	76.3	39.4	2.35	5	0.637	1.72	3.687
67	90	79.3	43.8	2.36	5	0.635	1.72	3.788
68	90	74.6	51.9	2.36	5	0.64	1.68	3.211
69	90	77.9	43.1	2.36	5	0.631	1.68	3.341
70	90	80.1	45.9	2.36	5.2	0.581	1.68	3.495
71	100	80.1	48	2.36	5.2	0.574	1.68	3.574
72	100	80.2	47.4	2.36	5.2	0.601	1.68	3.588
73	100	75.4	41.8	2.36	5.2	0.581	1.68	3.645
74	100	79.5	42.3	2.36	5.2	0.601	1.68	3.681
75	100	73.8	47.3	2.36	5.2	0.636	1.83	3.718
76	100	74.1	40.7	2.36	5	0.624	1.83	3.752
77	100	76.5	43.7	2.36	5	0.575	1.83	3.814
78	100	69.8	46.6	2.38	5	0.612	1.78	3.244
79	100	70.3	42.9	2.38	5.2	0.617	1.78	3.395
80	100	75.4	43.7	2.38	5.2	0.613	1.78	3.471
81	85	72.1	45.4	2.38	5.2	0.6	1.78	3.582
82	85	70.7	47.2	2.38	4.8	0.63	1.78	3.614
83	85	74.3	40.5	2.38	4.8	0.52	1.78	3.654
84	85	75.4	43.4	2.38	4.8	0.625	1.78	3.728
85	85	80.3	45.2	2.38	4.8	0.614	1.78	3.742
86	79	79.3	41.1	2.38	4.8	0.59	1.78	3.769
87	79	74.3	43.9	2.38	4.8	0.83	1.78	3.81
88	79	75.4	48.4	2.38	5.2	0.69	1.8	3.822
89	79	80.3	45.9	2.38	5.2	0.633	1.8	3.847
90	79	79.3	48	2.38	5	0.585	1.8	3.921
91	79	74.6	48	2.38	5	0.601	1.8	3.784
92	76	77.9	46.1	2.38	5	0.613	1.8	3.793
93	76	80.1	49.7	2.35	5	0.6	1.8	3.822
94	76	80.1	47.8	2.35	5.2	0.605	1.85	3.841
95	95	80.2	48.2	2.35	5.2	0.624	1.85	3.867
96	95	76.1	47.1	2.35	5.2	0.605	1.85	3.905
97	95	80.2	44.3	2.35	5.2	0.576	1.85	3.914
98	95	80.4	43.8	2.35	5	0.635	1.85	3.922
99	95	80.2	43.7	2.35	5	0.584	1.85	3.945
100	95	79.4	45.4	2.35	5	0.617	1.85	3.961

References

[1] Lees, A. (2019). Tensar International. Understanding Geotechnical Settlement.

[2] Geoengineer, The International Information Center for Geotechnical Engineers, (2017). Soil Settlement.

[3] CMW Geosciences, (2017). Ground settlement and subsidence.

[4] Ali, F., Al-samaraee, E. A. S. M. (2013). Field behavior and numerical simulation of coastal bund on soft marine clay loaded to failure. Electronic J. of Geotech. Eng. 18, 4027–4042.

[5] Lat, D. C., Ali, N., Jais, B., Baharom, B., Yunus, N. Z. M., Salleh, S. M., & Azmi, N. A. C. (2018). Compressibility characteristics of Sabak Bernam Marine Clay. MS&E, 342(1), 012082.

[6] Wan Yaacob W Z, Samsudin A R, Ramziemran M and Loon C Y. (2004). Natural Sorption Capability of Heavy Metals Granitic Residual Soil from Broga and Marine Clay from Sg.Besar. Geological Society of Malaysia, Bulletin 48 June 2004, p. 13-16.

[7] Suzuki K, Artificial neural networks: methodological advances and biomedical applications, BoD–Books on Demand, (2011)

[8] Raicea, R. (2017). Want to know how Deep Learning works? Here's a quick guide for everyone.

[9] Wei, F., Yao, G., Yang, Y., & Sun, Y. (2019). Instance-level recognition and quantification for concrete surface bughole based on deep learning. Automation in Construction, 107, 102920.

[10] Flores, A. C., Icoy, R. I., Peña, C. F., & Gorro, K. D. (2018). An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set. Paper presented at the 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST).

[11] Murthy, A. K., & Suresha. (2015). XML URL Classification Based on their Semantic Structure Orientation for Web Mining Applications. Procedia Computer Science, 46, 143-150. doi: https://doi.org/10.1016/j.procs.2015.02.005

[12] Torresani, L., & Lee, K.-c. (2007). Large margin component analysis. Paper presented at the Advances in neural information processing systems.