The 2020 World Congress on
Advances in Civil, Environmental, & Materials Research (ACEM20)
25-28, August, 2020, GECE, Seoul, Korea

Identifying pipejacking data patterns to assess ground conditions using support vector machine

Xue-Dong Bai[1)] and *Wen-Chieh Cheng[2)]

[1), 2)] *School of Civil Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China*
[2)] *w-c.cheng@xauat.edu.cn*

ABSTRACT

Detecting sudden change in geology (e.g., karst cavern and fault zone) is not an easy task. The change can result in jamming of shield machine or even geo-hazard (e.g., water ingress and surface subsidence) during tunnel excavation. Pipejacking parameters that relate closely to the surrounding geology have proliferated in recent years and present a substantial opportunity for the application of data-driven artificial intelligent (AI) techniques that can accentuate patterns in data from a dataset without reference to known, or labelled, outcomes. This study examined the potential for AI techniques to provide an ability of identifying the type of soils encountered during tunnelling process, thereby reducing the possibility of jamming and potential of geo-hazard. Further, a selection of the most popular parameter optimisation algorithms proposed in the literature was conducted for enhancing the accuracy of prediction. Their relative merits were evaluated by comparisons with a pipejacking case history undertaken in gravel and clayey gravel soils. The results highlighted an exciting potential for the use of support vector machine with optimization algorithms to identify the type of soils encountered during pipejacking.

1. INTRODUCTION

Pipejacking is an increasingly popular alternative to open-cut construction, especially in densely populated urban areas, because of the more efficient construction processes and reduced environmental impact. Identifying the type of soils encountered during pipejacking not only reduces the potential of geo-hazard but also prevents unplanned downtimes and operation costs. While geo-hazard prevention techniques exist, there remains significant motivation in the industry to further develop geological identification during tunnelling. In pipejacking, tunnelling parameters such as jacking force, cutter wheel torque, flow rate of feedline, pressure in slurry circulating system, and slurry density are highly variable due to their dependence on a number of

---

[1)] PhD Student
[2)] Professor

influencing factors including surrounding geology, lubrication, work stoppage and pipe deviation. For instance, the total jacking loads F_T required to advance a shield consists of the face resistance F_0 and the frictional resistance F_s. A significant body of research has indicated that effective lubrication can largely reduce F_s, whereas work stoppages and pipe deviation can cause significant and transient increases in F_s. Conversely, jacking into clayey gravel from gravel can cause an increase in F_0 due to increased contact of the shield face and therefore an increase in F_T. Although previous studies performed over the past three decades have significantly enhanced our understanding of the influencing factors and their influence on F_T, the relationship between surrounding geology and tunnelling parameters remains unclear. Systematic research to reveal the surrounding geology-tunnelling parameters relationship is therefore essential. In particular, identifying the response of tunnelling parameters to sudden change of geological condition is crucial for shield operators to adopt appropriate and timely countermeasures.

Traditionally, the manipulation of tunnelling parameters during pipejacking is highly dependent on the shield operator's accumulated site experience, yet the effectiveness of this process determines the safety of tunnel construction and adjacent properties. The proliferation of tunnelling parameters retrieved from modern tunnel shields presents substantial opportunity for the application of data-driven artificial intelligent (AI) techniques that can identify patterns in data without reference to known labels. This study examines the potential for data-driven AI techniques to identify the type of encountered soils, reducing the possibility of jamming and potential of geo-hazard. A selection of the most popular AI techniques proposed in the literature were considered for this purpose. Their relative merits were assessed by comparing AI predictions to monitored data from a recent case history of pipejacking in gravel and clayey gravel soils.

2. METHODOLOGY

2.1 Overview

Data-driven approaches identify characteristics of the measured system by utilising information retrieved from the measured data, rather than by modelling the system response. However, in most practical problems the measured data (i.e. the outputs) are not labelled. For this reason, 'unsupervised' machine learning algorithms, used to infer patterns in data without reference to known outcomes, is popular. The aim of this study is to develop an improved understanding of existing pipejacking parameters and their relationship with known geological changes during tunnelling, in which case 'supervised' machine learning is the optimal technique. Some popular examples of supervised machine learning algorithms include Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), and Support Vector Machines (SVMs). Although MARS generates a flexible model that can handle both linear and nonlinear relationships, it is less accurate for sparse data. RF, first proposed by Breiman (2001), is a 'tree-based' method. Each tree learns from its predecessors and updates the residual errors successively. While RF is a highly flexible technique it is susceptible to overfitting, particularly for small datasets. In light of these limitations, SVM is adopted for this study. The non-parametric nature of SVMs means that model

complexity is not influenced by the number of features (inputs) and they are therefore well-conditioned for high-dimensional datasets. Further, the use of kernels allows this technique to capture complex input-output mapping.

2.2 Support vector machines

SVMs can create a non-linear decision boundary by mapping the data through linear or non-linear kernels to a space with a high dimension (i.e. the 'feature space'). This means that data points which cannot be separated using a straight line in their original input space are 'lifted' to a new feature space where a straight hyperplane can separate the data into different classes. The hyperplane, mapped back to the original input space, therefore forms a non-linear curve and is represented by Eq. (1).

$$\boldsymbol{w} \cdot x_i + b \geq 1 \quad if\ y_i = 1$$

$$\boldsymbol{w} \cdot x_i + b \geq -1 \quad if\ y_i = -1 \tag{1}$$

where w is an adaptive weight vector, x is an input vector, y_i is the associated labels, and b is the bias. The hyperplane determines the margin between the classes; all the data points for the class '-1' are on one side, and all the data points for class '1' on the other. The distance from the closest point from each class to the hyperplane is equal. Thus, the constructed hyperplane searches for the maximum margin (termed the 'separating power') between classes. To prevent the SVM classifier from over-fitting noisy data (or to create a 'soft margin'), slack variables ξ_i are introduced here, allowing some data points to lie within the margin and the constant C defines the trade-off between the number of misclassification in the training data and margin maximisation. The objective function of the SVM classifier is the following minimisation formulation:

$$\text{Minimise } \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{n}\xi_i, \quad i = 1,2,\dots,n$$

$$\text{Subject to } y_i(w^T \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1,2,\dots,n \tag{2}$$

where C is the nonnegative penalty constant, n is the number of observations and ξ_i is a slack variable. When this minimisation problem is solved using Lagrange multipliers, the decision function (classification rule) for a data point x then becomes:

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x_j) + b \tag{3}$$

where α_i is a Lagrange multiplier. Every $\alpha_i > 0$ is weighed in the decision function and thus 'support' the machine. As SVMs are sparse, there will be relatively few Lagrange multipliers with a non-zero value. Since the outcome of the decision function only relies on the dot-product of the vectors in the feature space, it is not necessary to perform an explicit mapping to that space. Provided a kernel function K generates the same results, it can be used instead. Although the feature space can be of unlimited dimension (leading to a complex hyperplane), this unnecessary complexity is avoided here.

Popular selections for the kernel include linear, polynomial and sigmoidal functions; arguably the most common kernel function, the Gaussian Radial Basis Function (RBF), is also used here:

$$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\gamma^2}\right) \quad (4)$$

where γ is a kernel parameter, which is the width of the RBF, and $||x_i - x_j||$ is the dissimilarity measure. In general, the value of γ varies from 0 to 1.

2.3 Feature selection

A typical jacking force-distance plot is non-stationary and is unsuitable for direct application of SVMs. Further, the use of stationary data is desirable to accentuate patterns in the data. To this end, decomposition procedures can be used to disaggregate time series data into (stationary) feature-based sub-series where a weighted moving average dominates data features retained. Decomposition techniques were first developed by Persons (1919) to isolate salient features of a dataset. One of the most popular decomposition techniques is seasonal-trend decomposition using Loess smoothing (STL; Cleveland et al. 1990) which partitions the global series into three additive components as follows:

$$y_t = P_t + T_t + R_t \quad (5)$$

where P_t, T_t and R_t represent the periodic, trend and residual components respectively. Decomposition of the data involves sequential application of a Loess smoother. In this work, no significant periodic component was identified. Four feature variables (namely the trend and residual components of the measured torque and the total jacking load) were considered here for the application of SVMs. The pipejacking data decomposed into trend and residual components (i.e. without the periodic component) also limited false classifications. The weighted moving average of 3 m was adopted here not only to extract both components, but also to eliminate noise towards accentuating data features. In this paper, the current supplied to the cutterwheel is used as a proxy for torque and therefore the residual and trend torque components are plotted in 'Amps'. Since the variation in cutterwheel torque is more distinct for tunnelling in gravel (sand) than in other soils, its trend and residual components were used to identify 'gravel'. Pellet-Beacour and Kastner (2002) reported that local variations of total jacking load are generally linked to the varying face resistance. While jacking into fine soil governed gravel or sand layer (e.g. clayey gravel), it is possible for the cutting discs to sink into the cutting face, leading to full cutterhead-ground contact and producing variable face resistance. The trend and residual components of the total jacking load were thus used to detect 'clayey gravel'.

3. IMPLEMENTATION

The Support Vector Machine (SVM) algorithm discussed above was implemented using the Python module *Scikit-learn* (Pedregosa et al. 2011). All data were

preprocessed to maximise the efficiency and performance of the learning process and to ensure that the importance of the input dataset is equalised. A 'min−max scaler' was introduced to scale the dataset (Masters 1993), so that all data were laid between our specified range of minimum and maximum. The min-max scaler transforms the features of the input dataset to lie in the interval from 0 to 1. Given a set of input data x_1, x_2,…x_n, the scaled dataset z_1, z_2,…z_n will be:

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (6)$$

where min and max are our specified minimum and maximum values of the range to scale. The SVM used herein has two interdependent hyperparameters, namely *C* and *γ*. Designing an effective classifier requires the SVM parameters to be configured properly in advance. In this study, the SVM parameters were optimised using (a) genetic algorithm (GA), (b) particle swarm optimisation (PSO), and (c) grid search (GS). Genetic algorithm (GA) is a search algorithm which initially aims to simulate mechanisms of population genetics and natural rules of survival in pursuit of the ideas of adaptation. A typical GA begins with an initial set of random solutions, called a population. Each member within the population is termed a 'chromosome'. New chromosomes ('offspring') are created either by merging two chromosomes (from the current generation) using a crossover operator or by modifying a chromosome by means of a mutation operator. The parents and offspring that are selected based upon their objective values are key components of a new generation.

Particle swarm optimisation (PSO) was initially developed by Kennedy and Eberhart (1995) to simulate the graceful choreography of birds. In PSO, each solution represents a 'bird' in the flock and is referred to as a 'particle'. A particle is similar to a chromosome in GAs. Unlike GAs, the birds evolve their social behaviour, and accordingly their movement, towards a particular destination. The process therefore physically mimics a flock of birds when they fly. Each bird looks in a specific direction and they identify the bird that is in the best location when communicating together. Accordingly, each bird flies towards the best bird using a velocity that depends on its current position. Each bird then investigates the search space from its new local position. The process repeats until the flock reaches a desired destination. The birds therefore learn not only from their own experience (local search) but also from the experience of others (global search).

Grid search (GS) is an exhaustive search based upon a defined subset of the hyperparameter space. The hyperparameters are defined via 'lower bound', 'upper bound' and 'number of steps'. The data is first split into k subsets and in most instances the value of k is set to 10. One subset is used as a testing data and evaluated via the remaining k-1 training subsets. Various combinations of hyperparameters are trialled and the one with the best cross-validation accuracy is selected and used to train a GS-SVM model on the whole dataset.

4. PIPEJACKING DRIVES

4.1 Project overview

Cheng (2017, 2018, 2019a,b) describe a total of four drives in the soft alluvial

deposits of the Shulin district in Taipei County, Taiwan. One out of the four drives namely drives C was considered here to assess the selected AI techniques, . The length for Drives C was 75 m. Overburden depth relative to the tunnel crown for Drives C was measured at 10.8 m. The pipejacking was undertaken using a slurry shield machine with a 1.5 m diameter cutterhead. The trailing concrete pipe of 1.44 m in diameter and 1 m in length ensured a 30 mm overcut was created in the annulus area. The self-weight of each pipe was 12.6 kN. A highly viscous lubricant with Marsh cone viscosity of 38 mins was injected into the overcut annulus, thereby reducing friction resistance to viscous resistance.

4.2 Engineering geology

Fig. 1 presents the geological profile as determined from four geological boreholes (BH1 – BH4) installed close to drives C. The phreatic surface was located at approximately 4.5 m below ordnance datum (BOD). Fig. 2 shows the soil properties profile as determined from both field and laboratory tests. In addition, Drives C was rammed in a ground predominantly composed of gravel and clayey gravel (Fig. 3). Additional information on the project is available in Cheng et al. (2017, 2018, 2019a,b).

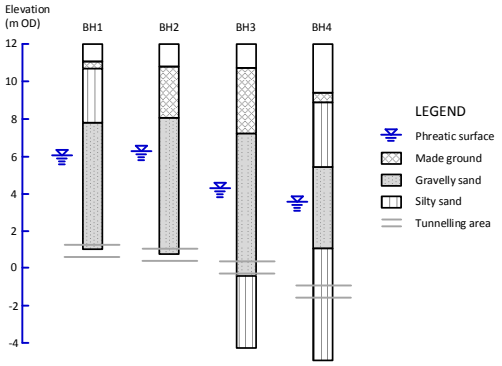

Fig. 1 Geological profile along the tunnel alignment

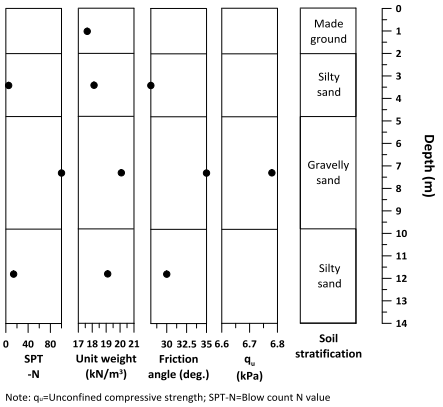

Fig. 2 Soil properties profile

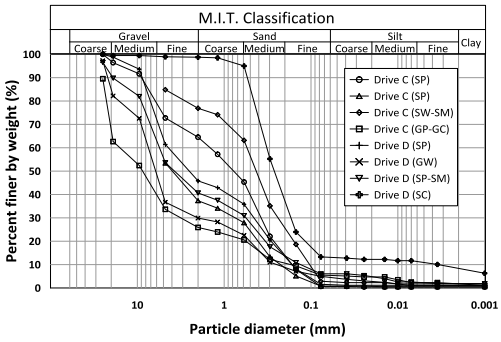

Fig. 3 Grain-size distribution curves for drive C

5. RESULTS AND DISCUSSION

5.1 Classification results

The three hyperparameter optimisation algorithms were first considered to evaluate their performance for this problem. To assess the feasibility of the optimisation algorithms, the fitness at each generation was traced, as shown in Fig. 4.

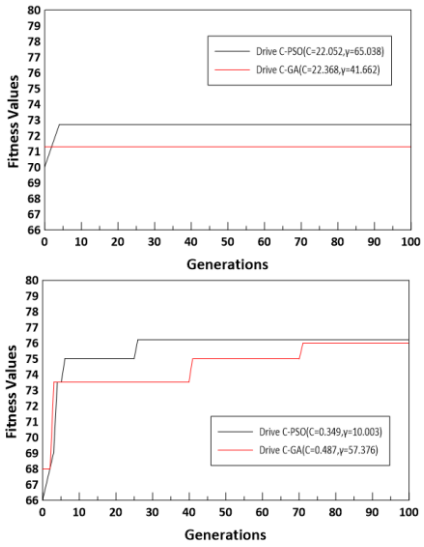

Fig. 4 Optimised results of the PSO and GA algorithms: (a) gravel identification and (b) clayey gravel identification

It can be observed that for gravel identification at drive C (Fig. 4a), the GA achieved an optimal solution immediately whereas for PSO it was achieved within four generations. For the identification of clayey gravel at drive C (Fig. 4b), GA required 71 generations to converge to a final solution whereas PSO required 26. In general, the PSO approach achieved faster convergence than GA. The wider the parameter range is, the more possibilities GS has of finding the best combination parameter. Notwithstanding that, GS is extremely time consuming especially when the number of possible different combinations of variables is rather high. Therefore, a grid search space ranging from 0.001 to 100 which is divided by a multiple of 10 was adopted here

to tackle the indicated issue, hence sacrificing the accuracy of prediction. The results of the parameter optimisation are tabulated in Table 1.

Table 1 Results of the parameter optimisation: (a) gravel identification and (b) clayey gravel identification

Hyperparameters used in identifying gravel at drive C	GA	PSO	GS
C	22.368	22.052	100
γ	41.662	65.038	100

Hyperparameters used in identifying clayey gravel at drive C	GA	PSO	GS
C	0.487	0.349	1
γ	57.376	10.003	1

The classification results present henceforth correspond to the optimised hyperparameters. We therefore use the more general terminology, 'SVM classifier', in the following sections. Further, we only discuss the cases that failed to correctly class the soil i.e. 'FN' in the residual-trend torque plot and 'FP' in the residual-trend jacking force plot. The results of the SVM classifier applied to the transformed data for drive C are shown in Fig. 5.

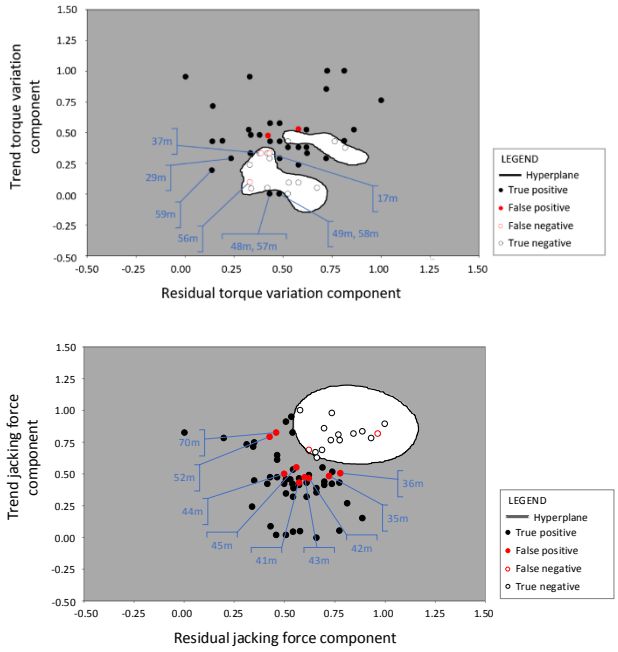

Fig. 5 Performance of the PSO-SVM model applied to drive C mapped to transformed parameter space: (a) identification of gravel (*C*=22.052, *γ*=65.038) and (b) identification of clayey gravel (*C*=0.349, *γ*=10.003)

Fig. 6 provides the reader with useful context in relation to the mapping of the TP, TN, FP and FN results from the transformed feature space back to the original (raw) parameter space. The SVM classifier provided excellent predictions with three FNs (see Fig. 5a). In this case, the FNs indicate that the gravels were erroneously classed as clayey gravels. Their locations in the original space are at jacked distances of 17 m, 37 m, and 56 m respectively. In contrast, several FPs appear in the predictions presented in Fig. 5b. FPs denote that the clayey gravels were erroneously identified as gravels. They appear at jacked distances of 35-36 m, 41-45 m, 52 m, and 70 m respectively.

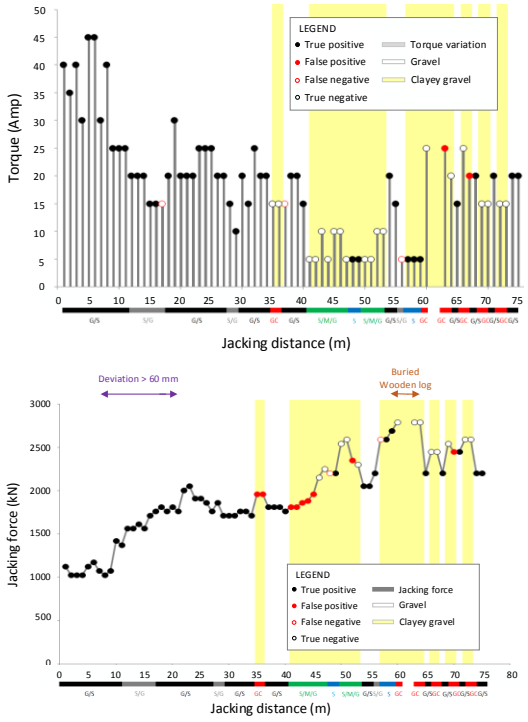

Fig. 6 Performance of the PSO-SVM model applied to drive C remapped back to initial parameter space: (a) identification of gravel (C=22.052, γ=65.038) and (b) identification of clayey gravel (C=0.349, γ=10.003)

5.2 Discussion

Most of the gravels present at drive C were successfully classed (Figs. 5a and 6a). The main cause to lead to the formation of the three FNs could be due to the fact that the occasional presence of sands, reported by Cheng et al. (2017), reduced their trend components Tt to nearly identical to or below 0.33 (16 Amp) (T_t > 0.33 can be classed as gravel) towards causing the misclassifications. Further, the clayey gravel is not frequently encountered along the tunnel alignment, which caused difficulties for the SVM classifier to define the hyperplane. However, the gravels at 29 m, 48-49 m, and 57-59 m jacking distances were correctly classed. The datapoints do not exactly correspond to gravels but sands with T_t below 0.33. The reason to explain why the

gravels can be correctly classed is that the lack of clayey gravel datapoints caused some difficulty in defining the hyperplane and the SVM classifier defined the noncontinuous boundaries by bypassing the sand datapoints in order to pursue improved predictive performance.

In contrast, nine FPs appear in the predictions in relation to identification of clayey gravel (Figs. 5b and 6b). A clayey gravel can be identified by satisfying two conditions; that are, the trend component $T_t > 0.58$ (2052 kN) and the residue component $R_t > 0.58$ (-5 kN). In spite that the FPs at 35-36 m and 41-45 m jacking distances were featured with $R_t > 0.58$, their T_t were below 0.58, induced by the gravels surrounding. Although the FPs at 52 m and 70 m jacking distances were featured with $T_t > 0.58$, the misclassifications occurred because of their $R_t < 0.58$, induced by traversing into the gravel from the clayey gravel. It is noteworthy that the predictions were surprisingly not affected by the effect of pipe deviation being greater than the threshold of 60 mm between 8 m and 21 m jacking distance.

It can be concluded that the discovery rate (DR) regarding identification of gravel being 94.1% at drives C, hence verifying the applicability of the trend component T_t, decomposed from the variation in cutterwheel torque, to class the gravel. The sands at drive C and striking buried wooden log led to misleading interferences about classification of gravel (inconsistent two regions) , as shown in Fig. 5a. Some FNs (i.e. 29 m, 48-49 m, and 57-59 m jacking distances at drive C) would have appeared if the SVM classifier did not pursue improved predictive performance by bypassing the datapoints. DR in relation to identification of clayey gravel was 59% at drives C, which also indicated a fairly good ability for the SVM classifier to class the clayey gravel. The clayey gravel can be successfully identified by higher T_t and R_t, in accordance with the established hyperplanes. In addition to the FPs induced by the reduced T_t, FP can also be induced by the reduced R_t resulting from jacking into the gravel from the clayey gravel.

The relative merits of the three optimisation algorithms were evaluated. Here the DRs and FARs determined using the three hyperparameter optimisation algorithms are summarised in Table 2.

Table 2 Performance of the optimisation algorithms: (a) gravel identification and (b) clayey gravel identification

Drive	Gravel identification					
	Discovery rate, DR (%)			False alarm rate, FAR (%)		
	GS	GA	PSO	GS	GA	PSO
C	94.1	92.2	94.1	9.1	18.2	9.1

Drive	Clayey gravel identification					
	Discovery rate, DR (%)			False alarm rate, FAR (%)		
	GS	GA	PSO	GS	GA	PSO
C	50.0	50.0	59.0	7.8	3.9	3.9

The PSO algorithm outperformed the other two algorithms. Although the PSO algorithm is more prone to becoming ‘trapped’ in local optima, it provided the best

balance between exploration and exploitation tendencies. In contrast, the GS algorithm provided the worst performance. The present analyses revealed that while the GS optimization was robust for the identification of optimal hyperparameter combinations, the required computational time was excessive. It is therefore highly reliable but suitable for only low dimensional datasets.

6. CONCLUSIONS

This paper has examined the potential for the use of AI techniques to identify geological conditions encountered during pipejacking. A selection of the most popular parameter optimisation algorithms was considered to improve the accuracy and efficiency of the AI predictions, namely genetic algorithms, particle swarm optimisation and grid search. Based on the results and discussion, some main conclusions can be drawn as follows:

(1) Decomposition of the data was implemented to transform the cutterwheel torque-jacking distance relationship and the jacking force-jacking distance relationship into feature-based sub-series for direct application of the SVM classifier. The optimal features were found to be the trend and residual components of both the cutterwheel torque and the total jacking force; the trend component dominated the identification of gravel, whereas the trend component T_t and the residual component R_t controlled the identification of clayey gravel. Further, the particle swarm optimisation algorithm provided the best performance due to its best balance between exploration and exploitation tendencies.

(2) Clayey gravels were not encountered frequently during pipejacking at drive C which caused some difficulty in defining the hyperplane. In addition, the surrounding geology could have implications on Tt, leading to false negatives. Some datapoints with low T_t did not cause false negatives because the SVM classifier bypassed them in order to pursue improved predictive performance. Similarly, surrounding geology could also have implications on T_t and R_t producing false positives. Further selection of highly-discriminant features, while classing the silt and/or sand, is deemed necessary to improve the accuracy of prediction.

(3) The performance of the optimisation algorithms was assessed using four performance pressures, namely, TP, FN, FP and TN, using the DR and the FAR indices. The PSO algorithm performed the best amongst the three optimisation algorithms because it provided the best balance between exploration and exploitation tendencies. While the GS sacrified the accuracy of predictions with a grid of low fineness.

REFERENCES

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

Persons, W.M. (1919), *Indices of Business Conditions: An Index of General Business Conditions*, Harvard University Press.

Cleveland, R., Cleveland, W., McRae, J., Terpenning, I. (1990), “STL: a seasonal-trend decomposition procedure based on loess,” *Journal of Official Statistics*, **6**, 3-73.

Pellet-Beaucour, A.L. and Kastner, R. (2002), "Experimental and analytical study of friction forces during microtunneling operations," *Tunnelling and Underground Space Technology*, **17**(1), 83-97.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011), "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, **12**(Oct), 2825-2830.

Masters, T. (1993), *Practical neural network recipes in C++*, Academic Press, San Diego, CA.

Kennedy, J. and Eberhart, R. (1995), "Particle Swarm Optimization," *Proceedings of IEEE International Conference on Neural Networks*, **4**, 1942-1948.

Cheng, W.C., Ni, J.C., Shen, J.S.L. and Huang, H.W. (2017), "Investigation into factors affecting jacking force: a case study," *Proceedings of the Institution of Civil Engineers - Geotechnical Engineering*, **170**(4), 322-334.

Cheng, W.C., Ni, J.C., Arulrajah, A. and Huang, H.W. (2018), "A simple approach for characterising tunnel bore conditions based upon pipe-jacking data," *Tunnelling and Underground Space Technology*, **71**, 494-504.

Cheng, W.C., Ni, J.C., Huang, H.W., Shen, J.S. (2019a), "The use of tunnelling parameters and spoil characteristics to assess soil types: a case study from alluvial deposits at a pipejacking project site," *Bulletin of Engineering Geology and the Environment*, **78**(4), 2933-2942.

Cheng, W.C., Wang, L., Xue, Z.F., Ni, J.C., Rahman, M., Arulrajah, A. (2019b), "Lubrication performance of pipejacking in soft alluvial deposits,' *Tunnelling and Underground Space Technology*, **91**, 102991.