International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 7 Issue 7 , July-2017,
ISSN (O): 2249-3905, ISSN(P): 2349-6525 | Impact Factor: 7.196

Machine Learning Applications for Site Characterization Based on CPT Data

Yeruva Ramana Reddy

Engineering Manager & Department of Civil Engineering

Indian Institute of Technology, India

yramanareddyiit@gmail.com

Abstract—*The primary purpose of this research is to explore ML Applications for Site Characterization that are Driven by CPT Data. An effective and adaptable approach for characterizing a site's geotechnical, lithostratigraphic, and hydrogeological conditions is cone penetration testing (CPT). Soil behavior type categorization using CPT data currently has serious limitations, frequently limiting its use to a small scale [1]. Site characterization and geotechnical design are dependent on a thorough understanding of geotechnical categorization. Field tests like the cone penetration test (CPT) are widely used because they are a quicker and more cost-effective method of sample recovery and testing than other methods. The subsurface information at a project site might be difficult to discern because of the lack of particular understanding of the soil forming histories and/or other preceding geological and human events. In order to get the most accurate representation of the subsurface, geophysical methods need to strike a careful balance between the use of high-frequency signals and receiver separations that extend far enough into the planet [1]. The greater the frequency of the signal, the better the vertical resolution. In order to get a high level of spatial or lateral resolution, sensors must be placed near together. Advancements in technology have made it possible to deploy many more sensors in a shorter period of time, while simultaneously improving spatial resolution.*

Keywords: Site characterization, CPT Data, Soil heterogeneity, Machine learning, Artificial neural networks (ANNs).

I. INTRODUCTION

Over the last decade, the use of machine learning algorithms in numerous disciplines of study has grown in popularity due to the rising availability of high-quality datasets. In the early phases of a project, one of the most important engineering tasks is the grouping of soils based on their related qualities. At this point, the project's viability is not yet proven. As a result, investing money comes with a significant level of risk. Cost-optimized (soil) research investigations are used in order to minimize the financial impact of an unworkable project. Due to significant expenditures connected with subsurface exploration, this practice is frequently kept to a minimum as much as feasible at the moment. Recently, the Cone Penetration Test (CPT) has earned a lot of attention for its ability to investigate soil conditions while still being very inexpensive [1,2].

Site Characterization data generated by experiments and simulations is now growing beyond the capacity of civil engineers to handle. Data-driven strategies for the detection of patterns across numerous length and time scales and structure-property correlations are necessary in order to uncover these links. The field of materials science stands to benefit greatly from these data-driven techniques. A overview of ML applications for metallic material characterisation is presented here [3]. Processing and structure of materials may have a significant impact on the qualities and performance of produced parts. As a result, this research aims to find out whether machine learning approaches can be used to forecast material properties. Qualitative Site Characterizations like toughness, brittleness, and ductility are important to distinguish between materials and components. Time-consuming and costly testing on materials such as tensile, compression, and creep are used in industry. This makes it simpler to generate Site Characterization information when ML techniques are used [3,4].

However, geotechnical and geological data are only accessible from a few numbers of geographically dispersed sites and may have to be inferred from archival or scheduled site research data at other locations [4]. The inferred subsurface model is plagued by high uncertainty due to a lack of detailed information on the formation process of the geological bodies and an inadequate number of drill logs and in-situ test results. Geotechnical engineers have long had to contend with challenges arising from subsurface uncertainty and the consequences of such uncertainty on geotechnical design. The primary objective of this study is to investigate several Machine Learning Applications for Site Characterization that are supported by CPT Data.

II. RESEARCH PROBLEM

The main problem that will be solved by this paper is to discuss how Machine Learning may be used for Site Characterization Based on CPT Data. A range of factors, including soil formation and deposition, as well as technical and monetary restrictions, contribute to the complexity of site characterization. To get the most out of geotechnical design, these uncertainties must be handled. Engineers may benefit from statistical techniques in this situation by gaining a better understanding of soil parameters and the related uncertainties [5]. A geotechnical construction project's planning and design relies heavily on correct site-specific soil and rock data. Soil and rock strata under the surface have intrinsic variety and unpredictability since they are naturally formed. It is thus necessary to take into consideration these spatial differences in soil/rock layering and engineering features of each recognized layer when designing and constructing a geotechnical system that is either embedded or based on the subsurface soils.

III. LITERATURE REVIEW

A. Overview of ML for site characterization

If there is sufficient data and a data-driven method for rule discovery, ML makes it possible for a computer to find the physical laws that lead to the provided data without the need for human intervention. The computer is used in traditional computational procedures in order to implement a pre-programmed algorithm that is given by a human subject matter expert. In contrast, machine learning systems may learn the principles that underlie a dataset by analyzing a subset of that dataset and constructing a model to make predictions. This process is known as "supervised learning." Despite this, humans are still required to choose appropriate machine learning models, which should accurately depict the data, as well as perform manual (sub-)tasks in the areas of pre-processing and feature development [6].

Because of the abundance of data, machine learning (ML) models may be used to gain new insights and patterns in the data. As a result of its large size and dimensions, big data poses a number of significant computational and statistical hurdles, including memory shortages and scalability issues as well as issues with noise buildup, correlation, and measurement errors.

Big data and machine learning have a lot of potential in the subject of materials science, which is only getting started. Processing, structure, characteristics, and performance are all crucial in materials research and engineering. No one is certain, however [6,7], how these components are linked. The so-called process-structure-property-performance chain may be used to learn more about the interrelationships between these components using ML approaches. It is one of the key goals of the project to enable, accelerate, or simplify the discovery and development of innovative materials using high-performance computing, automation, and machine learning. The discovery and measurement of critical material characteristics at high throughput is another goal of such methodologies in the area of materials research.

In addition to datasets that have been gathered via experimentation, several studies also show that data mining that is based on simulations. It has been demonstrated that experiment- and simulation-based data mining, when combined with machine learning methods, offers remarkable prospects to allow highly reliable identification of basic interrelationships within materials for the purpose of characterisation and optimizing in a scale-bridging way [8].

B. Site characterisation

Understanding the geological, hydrologic, and engineering features of a site includes the soil, rock, groundwater, and man-modified conditions in the subsurface (e.g. utilities, buildings, mines, and tunnels) that might affect site conditions [10]. It also covers the evaluation of pollutants in terms of their geographical and temporal distribution. A number of other names for this procedure have been coined, such as site inquiry, evaluation of the site, and characterisation of the site. The majority of project failures may be traced back to a lack of knowledge about the site circumstances that may have an influence on the project. Focusing on the geological and hydrologic conditions might have prevented these failures [10].

Site characterization is primarily concerned with predicting in situ soil parameters at each half-space point on a site using just a small number of tests. Geotechnical site investigation data is analyzed and interpreted in the field of site characterization. In order to define the spatial distribution of rock head heights, the neural network model necessitates the input of survey point coordinates (x,y) and surface elevation[10]. Using a contour map, we evaluated the trained network's ability to accurately predict the heights of rock heads across the study region. An excellent match is found between the neural network model's output and comparable contour maps produced by kriging. The capacity of this neural network-based technique to develop patterns or associations by training directly on the data is the primary benefit of this method. This eliminates the need to construct any difficult mathematical models and assumes that spatial changes would be the same everywhere. Researchers have employed neural networks to map the variation in permeability in order to determine landfill borders that will be built on a real site for the aim of characterizing applications linked to ground water characterisation. It was discovered that a simple technology called a neural network could correctly forecast the variation. Both the quality and amount of observations were examined as part of a study to see how accurate the suggested mapping process was. Using neural networks as a mapping tool, researchers found that they could better pinpoint the locations on a site where further subsurface investigation is needed[11].

C. ML approaches for Site Characterization

GIS and ANN have been used to create a multilayer perceptron, which has been demonstrated to be an effective method of describing the subsurface and determining aquifer parameters for ground-water flow modeling[11].

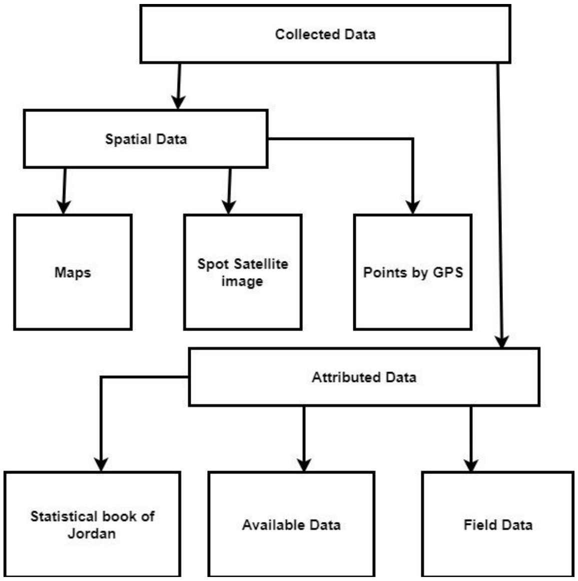

Fig i: A GIS representation

A leaking landfill's groundwater pollution may be accurately estimated utilizing a subsurface characterisation technique based on a modified counter propagation artificial neural network (ANN). The findings of this study demonstrate that it is possible to assess the level of contamination in the subsurface using a combination of structural equation modelling (used to decrease the complexity of the data), counter propagation artificial neural networks (ANN), and standard geostatistical approaches (kriging) [11]. This study highlights the capability of using ANN estimate (instead of kriging) to identify leachate polluted groundwater and assess water quality related with subsurface pollution at a full-scale site. Counter propagation ANN is shown to be a viable parameter estimation technique when several data sources are integrated to improve prediction accuracy and minimize uncertainty. These findings imply that The Support Vector Machine (SVM) was employed extensively in the development of a three-dimensional site characterisation model using Standard Penetration Tests. SVM parameters and have also been examined in detail. In this article, the findings obtained clearly show that the SVM is a powerful tool for characterizing the site [11,12]. In the work of Gomes and colleagues. High-resolution topographic data, computational simulations, and Bayesian analysis have been used to forecast the vertical reach of weathered material under soil-mantled steeper slopes to use a high-resolution spatial information. The efficiency and application of the DTB model and approach were shown in two case studies using synthetic and real-world regolith depth data [12]. Using root mean square error, it can be shown that the suggested DTB model with lumped parameters closely resembles the recorded regolith depth data (RMSE).

D. CPT Data Interpretation using machine learning

The interpretation of CPT data with the use of machine learning. In recent years, machine learning (ML) has emerged as a prominent method in a variety of scientific fields for the analysis of big datasets. The primary function fundamental difference between machine learning models and powerful computing techniques is that, instead of calculating the findings from an entry and a predetermined method, the ML model comes up with a solution through training from the information with the associated targets, independent of the exact method. This is in contrast to traditional computing strategies, which compute the findings by computing the data from an input and a specified solution (ANN, RF, etc.) [12]. In the current investigation, three distinct methods of machine learning—the Artificial Neural Network, Support Vector Machine, and Random Forest—are put to test and analyzed. The next part will provide a concise explanation of the function concepts that underlie their operation. Outlier identification and classification may be achieved using the Support Vector Machine (SVM) [13].

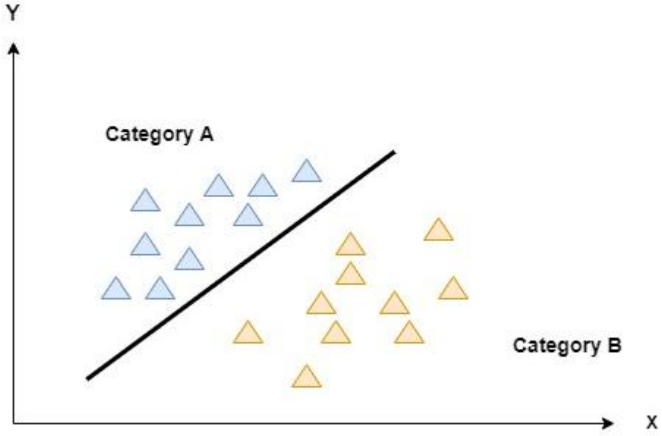

Fig ii: SVM Representation

Finding the differentiating hyperplanes in a higher or infinitely dimensional feature space that have the highest margin is one of the many jobs that an algorithm is tasked with doing. Some examples of these tasks include classification and regression. Generalization errors are reduced when the margin of error is big. An example of a linear support vector machine is shown in figure ii. Support vectors are the samples that are located on the margins of the graph. Recent applications of support vector machines (SVMs) in geotechnical engineering include the categorization of soil (13), the estimation of the bearing capacity of drilled piles based on CPT data (13,14), and the evaluation of soil compressibility.

The Artificial Neural Network is modeled after the way a real brain works, and it is composed of three distinct types of layers: first, there is the input layer, which is where the model receives the input variables; there is at least one hidden layer, which is where the data from the input nodes is applied in combination with the weights; and the output value, which is where the outcomes are quantified [14]. [Citation needed] In this particular investigation, a neural network was used that was trained via an iterative process called backpropagation. An error calculation and hidden layer are then performed by comparing simulated data to the data targets. This procedure is repeated until either an acceptable level of error is achieved or the rate of gradual improvements between trials is brought down to the minimum. ANNs have recently been utilized to categorize soils based on CPT data [3], determine soil properties [24], and predict cone resistance in a cone penetration test [14,15].

With CPT data, the ANN models are excellent at deducing soil behaviour patterns patterns. However, the predictions of soil types using grain-size dispersion was inadequate. The results that the RF models produced were the most accurate predictions for every possible set of input characteristics and target classes that were investigated. In addition, the favorable effects of new data on the pressures and soil physicochemical properties in the subsurface is evident in the enhanced model fit with the inclusion of the functional and total stress distribution and the hydraulic pore forces to the extracted features.

A collection of decision trees is what the Random Forest, abbreviated RF, is. A logistic regression is a kind of nonparametric supervised learning approach that may synthesize the classification methods derived from a set of data that includes attributes and tags, and then utilize the design of the forest to display such principles in solving regression and classification issues [15]. The answer provided by decision trees is understandable, and it is feasible to determine the weight that each input characteristic brings to the overall supervised learning techniques. Furthermore, geotechnical mathematical model [15], prediction of pile driveability [16], and most impressively, estimation of the unconfined compressive strength [15] have all been accomplished using random forest algorithms. . The decision process is supported in the first column of the nodes, and the Gini impurity, where measures the likelihood of a random source of data becoming incorrectly categorized and shows the importance of a split, is provided in the second line. The number of samples that have been seen in the node is shown on the third line of the graph. The ultimate categorization of the gathered data is shown in the fourth line of the table. The resultant class that was seen most often in the node [16,17] is shown in the very last line of the output.

IV. SIGNIFICANCE TO THE U.S

There are several reasons why site characterization is important in civil engineering projects in the United States. Remediation of a hazardous waste site would be incomplete without a thorough assessment of the area. Characterizing a site is essential for risk assessments, as well as for developing and executing remediation plans. Identification of contamination type and extent are the fundamental goals of site characterization [17]. Reliable geotechnical design parameters enable for more efficient and cost-effective transportation features since the dependability of the design is directly related to how well these design parameters are estimated. Designers, on the other hand, must design more conservatively to provide acceptable dependability when design parameters are less consistently established or when parameter values are uncertain; costs for building such systems are generally much larger than necessary with more reliable information.

It is simpler and less expensive to deal with project risks early in the project cycle when geotechnical hazards are detected. For instance, early detection of hazardous soil or rock may allow the relocation of roads or buildings to completely escape the danger or allow the selection of foundations for bridges or other structures to relieve hazards to the structures [18]. Due to delays in the planning phase or worse, after construction, expenditures for redesigning or remediating geotechnical hazards are unavoidably far higher than if the risks had been discovered earlier. First, if material records are available, they should be examined to see if they may help identify any pollutants that may be present at the site. Chemical-stock buying records, delivery records, storage records, and other documents might provide useful information about the sorts of pollutants that may be present at the site. The records of individual chemical accidents or leaks and waste disposal might also provide valuable information. Sadly, this information is sometimes insufficient, particularly for older or defunct websites.Inspections of probable sources of pollution are carried out after a review of available data. This involves looking for drums, leaky storage tanks, and abandoned disposal pits or injection wells. Once the source of contamination has been discovered, appropriate measures are taken to prevent further spread of the disease. Excavation and removal of porous medium, such as leaky tanks, is one example of this procedure.

V. FUTURE IN THE U.S.

Machine learning and CPT data will continue to be crucial to the process of site characterization in the future. In recent years, machine learning (ML) has emerged as a prominent method in a variety of scientific fields for the analysis of big datasets. The functional concept of machine learning models is fundamentally distinct from that of traditional computer algorithms in the sense that, rather than calculating the results from an input and a predetermined solution, the ML model discovers a remedy by studying from the input with the associated output (targets), irrespective of the particular method. This is the primary distinction between the two (ANN, RF, etc.). When considering the future, organizations that specialize in civil engineering in the United States will continue to employ the CPT as an essential part of their site inspection processes. Although the CPT is a good choice for the vast majority of deep foundations studies, the geology of Minnesota assures that rotary drilling and SPT will continue to be used on many projects, especially those that need ground improvement design [19]. Borings are also required for the installation of apparatus such as piezometers and inclinometers. "Targeted soil sampling and testing" based on preliminary CPT site investigation has only seldom been done, despite the fact that it has been debated for years. In the vast majority of instances pertaining to exploration, borings for projects are progressed via the use of a conventional sampling protocol, with samples being gathered at regular specified depths. This may be the effect of the borings being planned prior to the CPT work; but, in a broader sense, the extra cost of conducting the conventional sampling pattern is rather minimal when compared to the cost of collecting fewer samples after the rig has been moved to the site [19]. Targeted sampling has the potential to reduce the total number of samples required for laboratory work and boost productivity, but it also has the potential to provide a less comprehensive picture of the soil profile and fewer soil samples required for analyses such as moisture content and Atterberg limits. Usual sampling will probably remain the standard practice for all projects, with the exception of the most specialized ones (such as those involving landslides or temporary shoring that performs poorly), where certain samples could be obviously more significant for understanding the environment than others and where time may be a critical consideration. Despite the fact that CPT systems provide value, there is an understanding that two to three drill rigs are going to remain the fleet composition for the foreseeable future [19] despite the fact that excavation is slower and, in many instances, necessitated. When CPT rig was first purchased and deployed, local geotechnical consultants jumped at the opportunity to use the innovation and now provide their expertise to the state on architecture projects as well as any others that call for further geotechnical information. It is anticipated that there will be an increase in CPT usage and/or output as a result of the training of a modern trend of operators and the installation of better, more dependable equipment.

VI. CONCLUSION

Neural networks may be used to describe soil behavior under uniaxial strain situations, according to this study report. According to the results, site characterisation is an important step in deciding whether or not long-term repository performance is acceptable. Essentially this is a new way of doing things for an age-old method of engineering. It is possible to learn more about the subsurface via geophysical site characterisation. Geological features of underground areas are often sought for in certain situations. As a result, consultants and contractors have become increasingly familiar with CPT. With the absence of an in-house design guide for using the approach in many different kinds of projects, an obstacle to widespread adoption was highlighted (e.g., shallow foundation bearing and settlement). An understanding of the underlying stratigraphy or structure may be useful in the design of a treatment facility for contaminant contamination, for example. Geophysical technologies are used to get information about the subsurface while causing the least amount of damage to the environment. It is not uncommon for geophysical approaches to be used in conjunction with drill hole data in order to offer more comprehensive coverage. For the geophysical interpretation to be more accurate, the well data must be taken into consideration first. Despite this, it is one of a kind in a lot of ways. A new level of precision and accuracy is required in every aspect of field work, quality control, regulatory compliance, and public supervision. Considerably, the increasing the use machine learning and CPT on sites will enable the opportunity to gather significantly bigger quantities of high-quality data. This will allow for the development of precise profiles of soil strength and stiffness, as well as precise cross sections showing thin continuous layers. These factors will eventually have an influence on design choices and the cost of the project.

REFERENCES

[1] A. A. Basma and N. Kallas, "Modeling soil collapse by artificial neural networks," *Geotechnical and Geological Engineering*, vol. 22, no. 3, pp. 427–438, 2004. Available: https://doi.org/10.1023/b:gege.0000025044.72718.db

[2] B. Bhosale, "Curvelet Interaction with Artificial Neural Networks," in *Artificial Neural Network Modelling*. Cham: Springer International Publishing, 2016, pp. 109–125. Available: https://doi.org/10.1007/978-3-319-28495-8_6

[3] T. Lunne, *Cone penetration testing in geotechnical practice*. London: Blackie Academic & Professional, 1997.

[4] Y. Y. An and B. T. Wang, "Multifunctional Piezocone Penetration Testing in Geotechnical Practice," *Applied Mechanics and Materials*, vol. 90-93, pp. 250–254, Sep. 2011. Available: https://doi.org/10.4028/www.scientific.net/amm.90-93.250

[5] P. K. Robertson, "Soil classification using the cone penetration test," *Canadian Geotechnical Journal*, vol. 27, no. 1, pp. 151–158, Feb. 1990. Available: https://doi.org/10.1139/t90-014

[6] "Soil classification using the cone penetration test. Note," *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, vol. 27, no. 6, p. 348, Dec. 1990. Available: https://doi.org/10.1016/0148-9062(90)91219-w

[7] D.-K. Kim, "Investigations of Soil Classification Methods using Cone Test Results," *Journal of the Korea Academia-Industrial cooperation Society*, vol. 10, no. 7, pp. 1668–1672, Jul. 2009. Available: https://doi.org/10.5762/kais.2009.10.7.1668

[8] Z. Zhang and M. T. Tumay, "Statistical to Fuzzy Approach Toward CPT Soil Classification," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 125, no. 3, pp. 179–186, Mar. 1999. Available: https://doi.org/10.1061/(asce)1090-0241(1999)125:3(179)

[9] H. Yoon, P. Yoon, E. Lee, G.-B. Kim, and S.-H. Moon, "Application of machine learning technique-based time series models for prediction of groundwater level fluctuation to national groundwater monitoring network data," *Journal of the Geological Society of Korea*, vol. 52, no. 3, pp. 187–199, Jun. 2016. Available: https://doi.org/10.14770/jgsk.2016.52.3.187

[10] S. E. Cho, "Probabilistic stability analyses of slopes using the ANN-based response surface," *Computers and Geotechnics*, vol. 36, no. 5, pp. 787–797, Jun. 2009. Available: https://doi.org/10.1016/j.compgeo.2009.01.003

[11] N.-S. Park and S.-E. Cho, "Development of Seismic Fragility Curves for Slopes Using ANN-based Response Surface," *Journal of the Korean Geotechnical Society*, vol. 32, no. 11, pp. 31–42, Nov. 2016. Available: https://doi.org/10.7843/kgs.2016.32.11.31

[12] J.-S. Wang and 王俊翔, "Stratigraphic profiling and probabilistic site characterization based on cone penetration test," 學位論文 ; thesis, 2016. Available: http://ndltd.ncl.edu.tw/handle/53006556052024244811

[13] M. Lech, M. Bajda, and K. Markowska-Lech, "The use of resistivity and seismic cone penetration tests for site characterization," *Annals of Warsaw University of Life Sciences - SGGW. Land Reclamation*, vol. 40, no. 1, pp. 87–96, Jan. 2008. Available: https://doi.org/10.2478/v10060-008-0040-3

[14] A. Tillmann, A. Englert, Z. Nyari, I. Fejes, J. Vanderborght, and H. Vereecken, "Characterization of subsoil heterogeneity, estimation of grain size distribution and hydraulic conductivity at the Krauthausen test site using Cone Penetration Test," *Journal of Contaminant Hydrology*, vol. 95, no. 1-2, pp. 57–75, Jan. 2008. Available: https://doi.org/10.1016/j.jconhyd.2007.07.013

[15] M. Wazir, M. Herwan, S. Abd., and H. Shareef, "Hybrid Genetic Algorithm-Support Vector Machine Technique for Power Tracing in

Deregulated Power Systems," in *Real-World Applications of Genetic Algorithms*. InTech, 2012. Available: https://doi.org/10.5772/36196

[16] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 39, no. 1, pp. 43–62, Nov. 1997. Available: https://doi.org/10.1016/s0169-7439(97)00061-0

[17] B. Bhattacharya and D. P. Solomatine, "Machine learning in soil classification," *Neural Networks*, vol. 19, no. 2, pp. 186–195, Mar. 2006. Available: https://doi.org/10.1016/j.neunet.2006.01.005

[18] P. Vidyullatha, D. R. Rao, Y. Prasanth, R. Changala, and L. Narayana, "Integrating Different Machine Learning Techniques for Assessment and Forecasting of Data," in *Emerging Research in Computing, Information, Communication and Applications*. New Delhi: Springer India, 2015, pp. 123–130. Available: https://doi.org/10.1007/978-81-322-2553-9_12

[19] P. U. Kurup and E. P. Griffin, "Prediction of Soil Composition from CPT Data Using General Regression Neural Network," *Journal of Computing in Civil Engineering*, vol. 20, no. 4, pp. 281–289, Jul. 2006. Available: https://doi.org/10.1061/(asce)0887-3801(2006)20:4(281)