Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-1330

A machine learning model for predicting the unconfined compressive strength of MICP treated

Muhammad Nouman Amjad Raja
Department of Civil and Environmental Engineering, United Arab Emirates University, Al-Ain, UAE.
nouman.raja@uaeu.ac.ae

Tarek Abdoun
New York University-Abu-Dhabi, Abu-Dhabi, UAE

Waleed El-Sekelly
New York University-Abu-Dhabi, Abu-Dhabi, UAE

ABSTRACT: This study utilizes a machine learning-based approach, namely, Multivariate Adaptive Regression Splines (MARS) model, to predict the unconfined compressive strength (UCS) of sands treated with Microbially Induced Carbonate Precipitation (MICP). The dataset includes experimental data with key predictors such as median grain size (D50), coefficient of uniformity (Cu), void ratio (e), concentrations of urea (Mu) and calcium (Mc), optical density (OD), and calcium carbonate content (Fca). The MARS model effectively captures the nonlinear relationships between these input variables and UCS while remaining interpretable through the use of basic functions. Performance metrics reveal that the model achieves a mean square error (MSE) of 0.299 and a coefficient of determination (R2) of 0.929 in testing phase, demonstrating its reliability in predicting UCS. The results underline the MARS model's potential as a powerful and interpretable tool for UCS prediction, providing practical insights for optimizing MICP treatment protocols. This study highlights the model's ability to balance accuracy and interpretability, making it a practical choice for applications in MICP technology

KEYWORDS: MARS, Machine learning, MICP treatment, Prediction.

1 INTRODUCTION

With rapid urbanization and population growth, stabilizing loose soil has become increasingly necessary for construction projects. One innovative approach to achieving this is Microbially Induced Carbonate Precipitation (MICP), a cutting-edge technique that enhances soil compressive strength to ensure structural loads are securely transferred to the underlying ground. This process involves bacteria that produce urease—an enzyme that breaks down urea, releasing carbonate ions. These ions interact with calcium ions in the environment, resulting in the formation of calcium carbonate. This calcium carbonate acts as a natural binder, solidifying soil particles and significantly improving the soil's properties (Liu et al., 2021; Al Qabany and Soga, 2013). Unlike traditional soil stabilization methods, MICP offers a sustainable and eco-friendly solution by positively impacting the surrounding ecosystem. Additionally, it delivers high efficiency, causing minimal environmental disruption, making it a highly effective and environmentally responsible choice for soil improvement. (Cheng et al., 2017; Choi et al., 2016).

Despite of several experimental studies, limited efforts were made to predict the UCS of MICP treated sands. Current predictions for UCS are largely limited to linear or curvilinear models that primarily consider calcium carbonate (CaCO3) as the sole influencing factor (Cheng and Cord-Ruwisch, 2014). However, many studies has indicated that various other parameters, such the properties of the sand and MICP solution and the level of bacterial activity, can also significantly impact the UCS of MICP treated sand (Naqeeb et al., 2024; Tang et al., 2020). To encourage the wider use of MICP treatment, it is crucial to gain a deeper understanding of the strength enhancement mechanisms by developing comprehensive prediction models can capture the subtle complex (nonlinear) relations among the multiple key influencing factors.

In the recent past, Machine learning (ML) models were developed to predict the UCS of MICP treated sands (Khoshdel Sangdeh et al., 2024; Naqeeb et al., 2024; Talamkhani, 2023; Wang and Yin, 2021). Although these models, including hybrid neural networks, gradient boosting machines, and random forests, have demonstrated success in prediction, they are often criticized for their opaque, 'black-box' nature. Only a few studies, such as those by Naqeeb et al. (2024) and Sangdeh et al. (2024) have constructed gene expression programming (GEP) based models to forecast the UCS of MICP treated sands. GEP stands out because it can be translated into explicit mathematical formulas. Nonetheless, the accuracy achieved during the testing/validation phases, with R^2 values of 0.87 and 0.77 respectively, suggests room for improvement. Therefore, there is a need to explore other ML models that offer both transparency and the ability to be expressed in mathematical terms.

Over the past decade, the multivariate adaptive regression splines (MARS) modeling approach has gained popularity across various research domains. Developed initially by Friedman (Friedman, 1991), MARS offers a non-destructive, economical way to identify non-linear associations between inputs and their outputs. Unlike other ML algorithms such as artificial neural networks (ANNs) or support vector regression (SVR), MARS is noted for its computational efficiency, straightforward interpretability, and resilience in handling data irregularities. Recently, its effectiveness has been demonstrated in numerous geotechnical and ground engineering scenarios. An excellent review highlighting the cutting-edge applications of MARS in geotechnical engineering is available in literature (Zhang, 2020). In this study, MARS model is developed and implemented to predict the UCS of MICP treated sands. For calibrating and validating the MARS model, the pertinent dataset has been developed using the historical data.

2 DATABASE ESTABLISHEMNT

For building MARS model, it is imperative to have insights regarding the inputs affecting the UCS of MICP treated sand. For this study, nine inputs, namely, diameter of sand at 50% passing (D_{50}), coefficient of uniformity of sand (Cu); void ratio (e); density (optical) of bacterial solution (OD); urea concentration (M_u); calcium concentration (M_c) and calcium carbonate content (F_{ca}) are considered to be most influential parameters in determining the UCS of MICP treated sand (Talamkhani, 2023; Wang and Yin, 2021). The dataset utilized is the 351 tests reported by various researchers (Cheng et al., 2017, 2013; Cheng and Cord-Ruwisch, 2014; Dejong et al., 2013; Mahawish et al., 2018; Mujah et al., 2019; Nafisi et al., 2020; Paassen et al., 2010; Wang et al., 2020; Wen et al., 2019; Xiao et al., 2019; Zhao et al., 2014) and available in compiled form in Wang and Lin (2021). Figure 1 shows the correlation matrices of complete dataset.

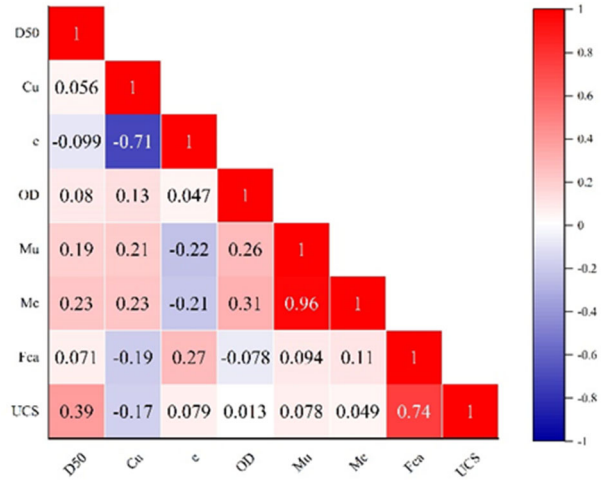

Figure 1. Correlation matrices of complete dataset.

3 MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

MARS is a flexible, nonlinear regression method based on a divide-and-conquer approach. It captures relationships between input variables and the response variable by splitting the training data into smaller, piecewise segments (splines) with varying slopes. These smooth polynomials, either linear or cubic, allow MARS to analyze multivariate data and quantify predictor contributions through basic functions (BFs) to simulate the response variable. Unlike traditional regression techniques, MARS does not rely on statistical assumptions for modeling relationships between predictors and responses (Friedman, 1991). The endpoints of the segments, known as knots, mark transitions between data regions. These knots enable the creation of piecewise linear or cubic curves (basis functions) that accommodate non-linearities, thresholds, and bends in the data (Zhang, 2020).

MARS uses stepwise searches to select BFs, while location of the knots are determined via adaptive recursive regression methods. The modeling process consists of two phases: a forward pass and a backward pass. In the forward pass, potential knots and functions are sequentially added until the residual error is minimized, resulting in an overfitted model (Raja and Shukla, 2021). The backward pass refines this model by removing less significant terms step by step.

Mathematically, after applying the linear combinations of BFs and their interactions, MARS can be represented as follows:

$$f(X) = \alpha_o + \sum_{n=1}^{m} \alpha_n BF(X) \quad (1)$$

where α_o is a constant estimated via least squares, and α_n represents BFs that may include individual splines or interactions between splines.

MARS follows a forward-backward approach to optimize the model. It begins with a single intercept α_o and iteratively adds BFs that significantly reduce the residual error, automatically incorporating interactions. The forward phase stops when error reduction is minimal or a predefined number of BFs is reached. In the backward phase, the model is pruned by removing less impactful BFs based on Generalized Cross Validation (GCV), a regularization metric that prevents overfitting.

For further mathematical details and examples of MARS applications, readers may refer to studies such as Friedman and Roosen (Friedman and Roosen, 1995). Figure 2 illustrates the MARS framework, employed in solving nonlinear multivariate prediction problem of UCS of MICP treated sands.

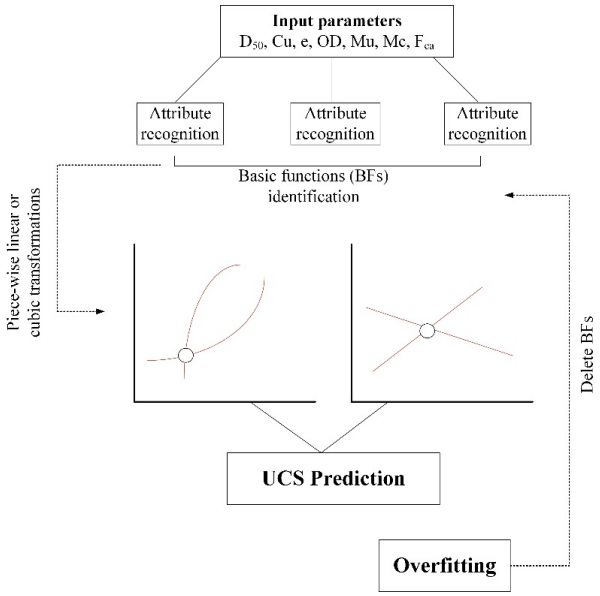

Figure 2. MARS architecture employed to predict UCS.

4 RESULTS AND DISCUSSIONS

For developing the MARS models, the complete dataset has been divided into 80% training and 20% testing data. Masters (Masters, 1993) recommended that the training and testing dataset statistical properties should be same for effective ML model training. Hence, several iterations were made to divide the data in a way that properties like minimum (min.), maximum (max.) and standard deviation (SD) of all the variables in training and testing dataset are close to each other (see Table 1). In spite of these attempts, there is a moderate discrepancies in some variables. This is due to the non-repeatability of certain values/scenarios in the dataset. Nonetheless, the whole dataset is kept and no point is excluded from the original database.

Table 1. Statistical properties of dataset.

Variables	Min. (Tr)	Max. (Tr)	SD. (Tr)	Min. (Ts)	Max. (Ts)	SD. (Ts)
D_{50}	0.12	1.6	0.30	0.14	1.6	0.34
C_u	1.17	6.23	1.14	1.25	6.23	1.12
e	0.43	1.04	0.07	0.43	0.98	0.07
OD	0.3	4.46	1.21	0.3	4.46	1.14
M_u	0.1	1.5	0.34	0.1	1.5	0.34
M_c	0.1	1.5	0.33	0.1	1.5	0.34
F_{ca}	1.49	29.47	6.51	1.76	27.16	6.46
UCS	0.05	14.23	1.97	0.07	10.59	1.99

Note: Tr for training dataset and Ts for testing dataset

The generalized recursive partitioning method was utilized to approximate the optimal function. To optimize knots, third-order interactions were analyzed with a degree of freedom (DoF) of three. A general guideline suggests that the maximum number of BFs should be two to four times the number of input parameters. Accordingly, with seven inputs, the maximum allowable BFs were set to 28. After initializing the dual-phase mechanism, the model with the lowest Generalized Cross-Validation (GCV) value was chosen as the optimal one. The variation of BFs with MSE for training and testing dataset are shown in Figure 3.

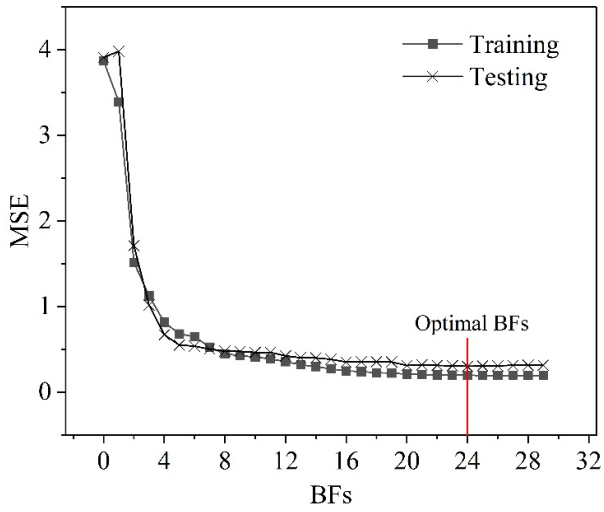

Figure 3. Variation of MSE versus BFs in training and testing data.

Table 2. Basic functions and their formula for MARS model.

Basic Functions	Formula	Basic Functions	Formula	Basic Functions	Formula
BF1	max(0, F_{ca} - 1.49)	BF9	max(0, 2.36 - OD) × max(0, 0.5 - M_c) × BF1;	BF17	max(0, e - 0.61) × BF19
BF2	max(0, D_{50} - 0.687) × BF1	BF10	max(0, M_u - 0.75) × BF3	BF18	max(0, e - 0.59) × BF1
BF3	max(0, C_u - 1.44) × max(0, 0.687 - D_{50}) × BF1	BF11	max(0, F_{ca} - 15.8) × BF9	BF19	max(0, 0.59 - e) × BF1
BF4	max(0, 1.44 – C_u) × max(0, 0.687 - D_{50}) × BF1	BF12	max(0, 15.8 - F_{ca}) × BF9	BF20	max(0, D_{50} - 0.352) × BF19
BF5	max(0, D_{50} - 0.25)	BF13	max(0, C_u - 2.21) × BF1	BF21	max(0, M_c - 0.75)
BF6	max(0, 0.25 - D_{50})	BF14	max(0, 2.21 - C_u) × BF1	BF22	max(0, 0.75 - M_c)
BF7	max(0, M_c - 0.5) × BF1	BF15	max(0, e - 0.66) × BF19	BF23	max(0, F_{ca} - 22.6)
BF8	max(0, OD - 2.36) × max(0, 0.5 – M_c) × BF1;	BF16	max(0, 0.66 - e) × BF19	BF24	max(0, M_u - 0.1)

$$UCS = -1.25314 - 0.734721 \times BF1 + 2.20061 \times BF2 + 4.28556 \times BF3 - 9.7011 \times B + 0.709033 \times BF5 - 25.0213 \times BF6 + 0.305691 \times BF7 + 0.712013 \times BF8 + 0.24218 \times BF9 - 1.42018 \times BF10 + 1.81117 \times BF11 + 1.71728 \times BF12 - 0.541241 \times BF13 + 0.969853 \times BF14 - 12.7306 \times BF15 + 15.6776 \times BF16 + 19.7355 \times BF17 - 6.21942 \times BF18 - 2.00968 \times BF19 - 1.98262 \times BF20 - 3.40505 \times BF21 + 1.04883 \times BF22 - 0.11333 \times BF23 + 2.02551 \times BF24 \quad (2)$$

In training, MARS model achieved the accuracy of 0.947 whereas for testing it is 0.929. Figure. 4 shows the scatter plot of the testing dataset along with the ideal prediction line (1:1).

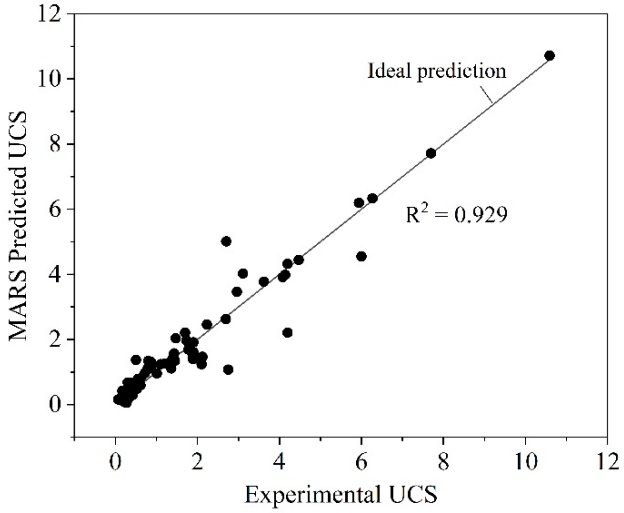

Figure 4. Scatter plot of the MARS model in testing phase.

Figure 5 represents the violin plots that helps visualize the probability density of the experimental and predicted data. It can be observed that the both in training and testing phase, MARS model predictions are reasonably aligned with the experimental results.

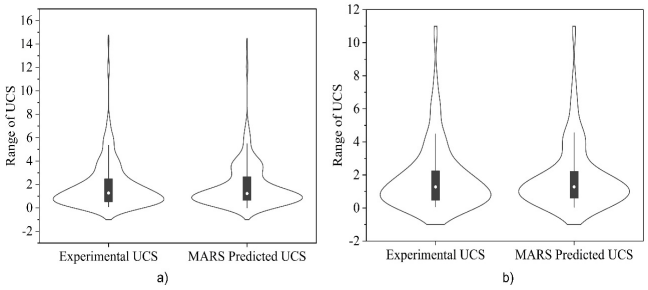

Figure 5. Violin plots for MARS in: a) training phase; b) testing phase.

The sensitivity analysis was also conducted via ANOVA decomposition technique. The results are shown in the Figure 6. Although, all the variables play an important role in predicting the strength of MICP treated sands, however, the highest relative score is achieved by F_{ca}, followed by D_{50} and e.

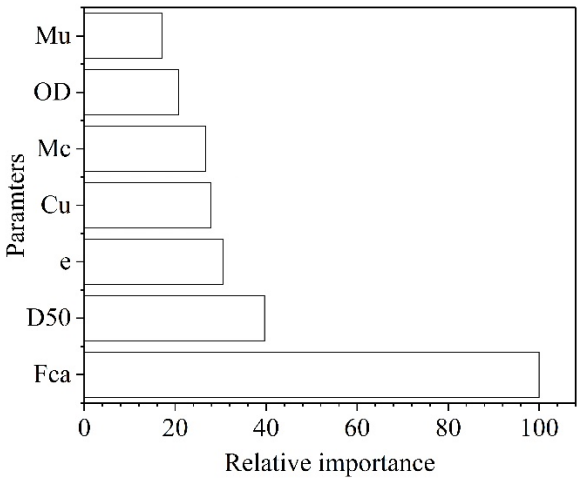

Figure 6. Relative importance of input variables.

5 CONCLUSION

In this study, an attempt is made to predict the UCS of MICP treated sands using MARS modeling technique. The historical experimental dataset was divided into 80% for training and 20% for testing. The results indicate that the MARS model predicted the UCS with reasonable accuracy, achieving R^2 values of 0.947 in the training phase and 0.929 in the testing phase. Furthermore, violin plots reveal that the model not only provides good accuracy but also effectively replicates the probability density in both datasets. Sensitivity analysis indicates that the variables F_{ca}, D_{50}, and e play the most important roles in predicting the UCS of MICP-treated sands. Notably, the model has been translated into a simple mathematical equation, enabling easy implementation by non-AI experts for quick estimation of UCS.

6 REFERENCES

Liu, J., Li, G. and Li, X. 2021. Geotechnical engineering properties of soils solidified by microbially induced $CaCO_3$ precipitation (MICP). *Advances in Civil Engineering*, 2021(1), 6683930.

Al Qabany, A. and Soga, K. 2013. Effect of chemical treatment used in MICP on engineering properties of cemented soils. *Geotechnique*, 63(4), 331–339.

Choi, S.G., Wang, K. and Chu, J. 2016. Properties of biocemented, fiber reinforced sand. *Construction and Building Materials*, 120, 623–629,

Cheng, L., Shahin, M.A. and Mujah, D. 2017. Influence of key environmental conditions on microbially induced cementation for soil stabilization. *Journal of Geotechnical and Geoenvironmental Engineering*, 143(1), 04016083

Cheng, L. and Cord-Ruwisch, R. 2014. Upscaling effects of soil improvement by microbially induced calcite precipitation by surface percolation. *Geomicrobiology Journal*, 31(5), 396–406.

Naqeeb, N.M., Yar Akhtar, A., Hassan, W., Hasnain Ayub Khan, M. and Muneeb Nawaz, M. 2024. Artificial intelligence-based prediction models of bio-treated sand strength for sustainable and green infrastructure applications. *Transportation Geotechnics*, 46, 101262.

Tang, C.S., Yin, L.Y., Jiang, N.J., Zhu, C., Zeng, H., Li, H. et al. 2020. Factors affecting the performance of microbial-induced carbonate precipitation (MICP) treated soil: a review. *Environmental Earth Sciences*, 79(5), 94.

Wang, H.L. and Yin, Z.Y. 2021. Unconfined compressive strength of bio-cemented sand: state-of-the-art review and MEP-MC-based model development. *Journal of Cleaner Production*, 315, 128205.

Khoshdel Sangdeh, M., Salimi, M., Hakimi Khansar, H., Dokaneh, M., Zanganeh Ranjbar, P. and Payan, M. et al. 2024. Predicting the precipitated calcium carbonate and unconfined compressive strength of bio-mediated sands through robust hybrid optimization algorithms. *Transportation Geotechnics*, 46, 101235.

Talamkhani, S. 2023. Machine learning-based prediction of unconfined compressive strength of sands treated by microbially-induced calcite precipitation (MICP): a gradient boosting approach and correlation analysis. *Advances in Civil Engineering*, 2023(1), 3692090.

Friedman, J.H. 1991. Multivariate adaptive regression splines. The *Annals of Statistics*, 19(1), 1–67.

Zhang, W. 2020. MARS applications in geotechnical engineering systems. Singapore: Springer Singapore.

Xiao, Y., He, X., Evans, T.M., Stuedlein, A.W. and Liu, H. 2019. Unconfined compressive and splitting tensile strength of basalt fiber–reinforced biocemented sand. *Journal of Geotechnical and Geoenvironmental Engineering*, 145(9), 04019048.

Zhao, Q., Li, L., Li, C., Li, M., Amini, F. and Zhang, H. 2014. Factors affecting improvement of engineering properties of MICP-treated soil catalyzed by bacteria and urease. *Journal of Materials in Civil Engineering*, 26(12), 04014094.

Dejong, J.T., Soga, K., Kavazanjian, E., Burns, S., Van Paassen, L.A. and Al Qabany, A. et al. 2013. Biogeochemical processes and geotechnical applications: progress, opportunities and challenges. *Geotechnique*, 63(4), 287–301.

Cheng, L., Cord-Ruwisch, R. and Shahin, M.A. 2013. Cementation of sand soil by microbially induced calcite precipitation at various degrees of saturation. Canadian Geotechnical Journal, 50(1), 81–90.

Mahawish, A., Bouazza, A. and Gates, W.P. 2018. Effect of particle size distribution on the bio-cementation of coarse aggregates. *Acta Geotechnica*, 13(4), 1019–1025.

Mujah, D., Cheng, L. and Shahin, M.A. 2019. Microstructural and geomechanical study on biocemented sand for optimization of MICP process. *Journal of Materials in Civil Engineering*, 31(4), 04019025.

Nafisi, A., Mocelin, D., Montoya, B.M. and Underwood, S. 2020. Tensile strength of sands treated with microbially induced carbonate precipitation. *Canadian Geotechnical Journal*, 57(10), 1611–1616.

Van Paassen, L.A., Van Loosdrecht, M.C.M., Pieron, M. and Mulder, A. et al. 2010. Strength and deformation of biologically cemented sandstone. *In: Rock Engineering in Difficult Ground Conditions - Soft Rocks and Karst: Proceedings of the Regional Symposium of the International Society for Rock Mechanics*, EUROCK 2009.

Wen, K., Li, Y., Liu, S., Bu, C. and Li, L. 2019. Development of an improved immersing method to enhance microbial induced calcite precipitation treated sandy soil through multiple treatments in low cementation media concentration. *Geotechnical and Geological Engineering*, 37(2), 1015–1027.

Wang, Y., Konstantinou, C., Soga, K., DeJong, J.T., Biscontin, G. and Kabla, A.J. 2020. Enhancing strength of MICP-treated sandy soils: from micro to macro scale. arXiv preprint.

Raja, M.N.A. and Shukla, S.K. 2021. Multivariate adaptive regression splines model for reinforced soil foundations. *Geosynthetics International*, 28(4), 368–390.

Friedman, J.H. and Roosen, C.B. 1995. An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research*, 4(3), 197–217.

Masters, T. 1993. Practical neural network recipes in C++. Massachusetts, USA: Morgan Kaufmann Publishers.