Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 204 (2022) 172–179

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry Sciences and Computer Science Innovation

A comparison of unsupervised and supervised machine learning algorithms to predict water pollutions

Nurnadiah Zamri[a]*, Mohammad Ammar Pairan[b], Wan Nur Amira Wan Azman[c], Siti Sabariah Abas[d], Lazim Abdullah[e], Syibrah Naim[f], Zamali Tarmudi[g], Miaomiao Gao[h]

[a,b,c,d]*Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Campus, Besut, Terengganu, Malaysia*
[e]*Management Science Research Group, Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia Management Science Research Group,*
[f]*Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom*
[g]*Universiti Technology MARA, Johor Branch, Segamat, Johor, Maaysia*
[h]*Faculty of Biology, Medicine and Health, School of Medical Siences, University of Manchester, Manchester, United Kingdom*

Abstract

Clean and safe water is vital for our lives and public health. In recent decades, population growth, agriculture, industries, and climate change have worsened freshwater resource depletion and clean water pollution. Several studies have focused on water pollutions risk simulation and prediction in the presence of pollution hotspots. However, the increase and complexity of big data caused by uncertain water quality parameters led to a new efficient algorithm to trace the most accurate pollution hotspots. Therefore, this study proposes to offer different algorithms and comparative studies using Machine Learning (ML) algorithms. Ten different most widely used algorithms, including unsupervised and supervised ML, will be employed to categorize the pollution hotspots for the Terengganu River. Besides, we also validate algorithms' accuracies by improving and changing each parameter in ML algorithms. Our results list all the accurate and efficient ML algorithms for the classification of river pollutions. These results help to facilitate river prediction using efficient and accurate algorithms in various water quality scenario.

Keywords: Machine Learning; Deep Learning; Water Pollutions; Terengganu River

* Corresponding author. Tel.: +609-6993027; fax: +609-6993215.
E-mail address: nadiahzamri@unisza.edu.my

1877-0509
10.1016/j.procs.2022.08.021

1. Introduction

Groundwater, which includes rivers, streams, dams, lakes, reservoirs, creeks, and wetlands, is an important resource providing the main source of clean and safe drinking water; to domestic households, industries and agriculture [1]. Despite its greatest importance in maintaining human life and other habitats, including aquatic plants and wildlife, ground water is immensely faced to pollution coming from uninhibited human activities of industrialization and agriculture. Based on the United Nations World Water Development Report 2015, developing countries left almost 90% of untreated sewage that goes straight into water sources [2]. Meanwhile, the statistics from United Nations Educational, Scientific, and Cultural Organization (UNESCO) showed that around 300 to 400 megatons of waste coming from the industry had been discharged into the water source bodies each year [3].

Malaysia is one of the countries also faced with the high uncontrolled treatment of sewage or discharge from manufacturing and agro-based industries [4]. Ammoniacal Nitrogen (NH_3-N), Biochemical Oxygen Demand (BOD), and Suspended Solids (SS) are the most crucial parameter that causes river pollution. Effluent and ineffective sewage treatment coming from manufacturing and agro-based industries can contribute to high BOD. Meanwhile, uncontrollable domestic sewage and animal farming can contribute to high NH_3-N. Besides, improper land clearing activities and earthworks can contribute to high SS [4]. Continuous water quality statistics on these river pollution hotspots are needed to categorize which areas are polluted and must be treated. Several previous studies have discussed on water quality assessment in Malaysia [5, 6, 7]. However, the increase and complexity of big data caused by uncertain water quality parameters led to a new efficient algorithm to trace the pollution hotspots

Therefore, this study proposes to offer different algorithms and comparative studies using unsupervised and supervised Machine Learning (ML) algorithms to efficiently trace the river pollution hotspots. Several studies have discussed water quality with ML algorithms [8, 9, 10]. However, as far from our knowledge, no detailed comparative study on the application of ML in river quality assessment datasets is found in the literature. Thus, ten different most widely used algorithms including unsupervised and supervised ML which are Hierarchical Clustering (HC), K-Nearest Neighbors (KNN), Support Vector Classifiers (SVC), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Naïve Beyes (NB), Multi-Layer Perceptron (MLP), Random Forest (RF), Decision Tree (DT) and K-Means (KM) will be employed to categorize the pollution hotspots in Terengganu River which is one of the river parts in Malaysia. Besides, we also validate algorithms' accuracies by improving and changing each parameter in ML algorithms. This process is crucial in every ML algorithm for avoiding underfitting and overfitting [11]. Thus, the algorithms will make overfit and underfit to spurious designs and patterns in the training data and not generalize [12]. Overall, this paper is discussed on the detail comparison and evaluation for each ML algorithm based on the river pollutions classification's accuracy. Highest accuracy is retrieved from the preprocessing strategies and tuning processes on each algorithm.

2. Materials and Methods

This paper focuses on two types of ML, which are unsupervised and supervised. In unsupervised learning includes HC, the goal is to identify group patterns within the dataset. Supervised learning has KNN, SVC, LR, LDA, NB, MLP, RF, DT, and KM; the goal is to predict the polluted hotspots in the Terengganu River. We divided Terengganu River data into two official parts: training and validation. A training dataset is used to calibrate the algorithm's parameters, and a validation dataset is utilized to evaluate algorithm performance [13]. We began by assessing an unsupervised ML method, HC, a given number of clusters are estimated by iteratively assigning data points from datasets. After the exact number of clusters are retrieved, supervised ML methods, KNN, SVC, LR, LDA, NB, MLP, RF, DT, and KM, are performed to predict water pollution hotspots.

2.1. Datasets

The datasets used to present measurement of water pollutions in Terengganu River, Malaysia [14]. These datasets contain 405 samples with 27 features for five different levels of water pollutions. Five different levels water pollutions include Very Clean, Clean, Slightly Polluted, Polluted, and Highly Polluted. All data are in positive and

negative integers where each value represents the characteristics of the water pollutions level that allow the learning process from unsupervised and supervised ML algorithms.

2.2. Preparing for Datasets

For the experiments, the information is divided into two parts. The first part matches to the features (*X*); includes all the water quality parameters, and the second part matches to the classes (*Y*); includes all the river pollution hotspots. The features compose a matrix of size *pxq*, and the classes are a vector of size *qx1*, where *p* is the number of river pollution hotspots area and *q* is the number of water quality parameters. Hidden patterns in the datasets are discovered using unsupervised ML to retrieve the most suitable number of pollutions' clusters. Next, using the same 405 river pollution datasets, we subdivided them into two subsets: 80% training and 20% validation. The training dataset is utilized to calibrate the supervised ML algorithm, and the validation dataset is used to hyperparameter tuning and measure the accuracy. K-fold validation with k=10 is set for the hyperparameters tuning to calculate the accuracy for each algorithm. Two preprocessing techniques were used to improve the statistical significance which are Standard Scaler and Principal Component Analysis (PCA). Based on the original datasets, we created them into four different types of datasets for the training and validation of each supervised ML algorithm; 1) original data stated as Raw; 2) performed a scaling process; Standard Scaler; 3) applied PCA using Raw data to reduce data dimensionality with a retained variance of 80%; 4) applied both Standard Scaler and PCA. The combination of these preprocessing datasets is to retrieve the best performance for each algorithm.

2.3. Significance Tests

A significance test is performed to determine the differences in accuracy between each algorithm and then decide whether it is significant. The difference between the observed and expected accuracies is computed under a normal distribution. The accuracy can be calculated using the number of correct test predictions x and the number of test instances N, as follows:

$$\text{Accuracy}_i = \frac{x}{N} \tag{1}$$

$$H_0 : \text{Accuracy}_i - \text{Accuracy}_j = 0 \tag{2}$$

$$H_1 : \text{Accuracy}_i - \text{Accuracy}_j \neq 0 \tag{3}$$

These significant tests allow to determine the accuracy of each algorithm after the tuning process and decide whether the necessary of the parameter tuning and the relevancy of each supervised ML algorithm are necessary.

2.4. Tools

The preprocessing using Standard Scaler and PCA transformation were executed using preprocessing modules and decomposition from Python scikit-learns [15]. All unsupervised and supervised algorithms were executed using scikit-learn libraries from the Python programming language. Matplotlib libraries were used to create the images [16].

3. Clustering Unsupervised ML

Hierarchical Clustering (HC) is an algorithm to group patterns in the dataset. It is essential to visualize the group patterns before evaluating the classification of the water pollution hotspots. HC with Ward's method is created to retrieve the most balanced group patterns using the Raw and PCA dataset. Fig. 1 shows there are 5 major group patterns at distance=0.2. Meanwhile, Fig. 2 shows number of river pollution hotspots for each group patterns composition.

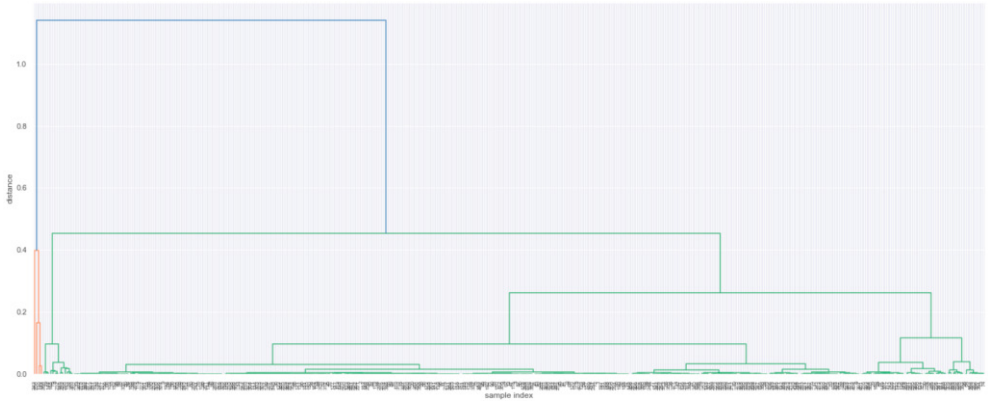


Fig. 1. HC using Ward's method

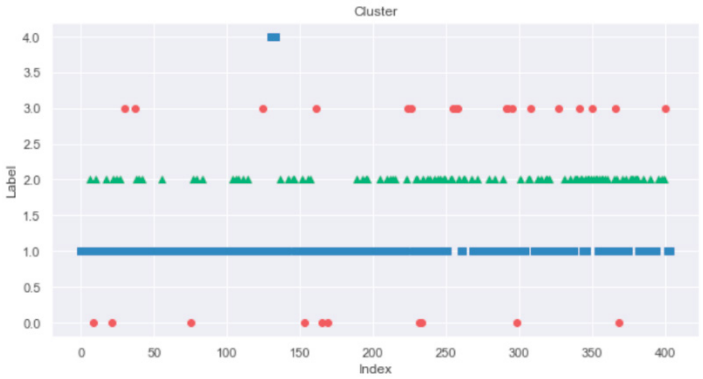


Fig. 2. Cluster composition

Based on Fig. 1 and Fig. 2 results, Table 1 listed all the river pollution hotspots for each class. These classes include classes from the original datasets, classes that contributed from Raw data, and classes that were retrieved by preprocessed data with PCA.

Table 1. Number of river pollution area for each class.

Class	Original datasets	Raw datasets	Preprocessed datasets with PCA
0	8	10	15
1	330	278	305
2	63	97	68
3	3	18	15
4	1	2	2

Table 1 shows there are some slights different in terms of the number of river pollution hotspots, between original datasets, Raw datasets, and preprocessed datasets with PCA. However, it is still in line with the five different levels of water pollution coming from the original datasets: Very Clean, Clean, Slightly Polluted, Polluted and Highly Polluted. These slights are different, maybe coming from the different types of datasets.

4. Predicting using Supervised ML

Parameters are used to tune the algorithms before running supervised ML. Four different parameters for four different algorithms are defined in Table 2 to discover the best behaviour using four types of datasets (Raw, Standard Scaler, PCA, and Standard Scaler+PCA). The best hyperparameters were calculated using different parameters and determined which dataset could be the most appropriate.

Table 2. Tuning parameters.

Tuning parameters	Detail description
Neighbors	Number of neighbors
C	Penalty parameter C of the error term.
Neurons	Number of neurons in hidden layers.
Clusters	Number of clusters.

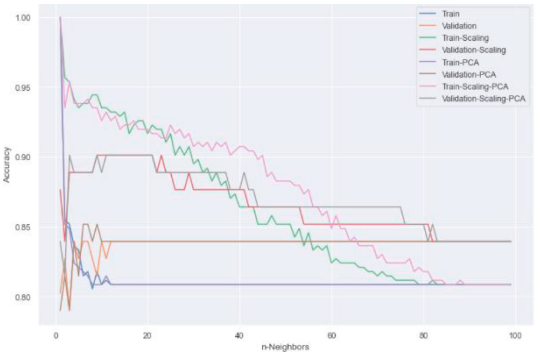
KNN, SVC, MLP, and KN used the parameters stated in Table 2 to tune the algorithms. Meanwhile, LDA is set as parameter default, RF and DT are tuned based on the number of estimators with the minimum and maximum sample splits. All these algorithms were trained and validated using four types of datasets. Figure 3 plotted the accuracy values of the training and validation processes on all types of datasets. RF and DT were not plotted since more than one hyperparameter was tuned. Meanwhile, the accuracy results of each algorithm are listed in Table 3.

Table 3. Accuracy for each algorithm.

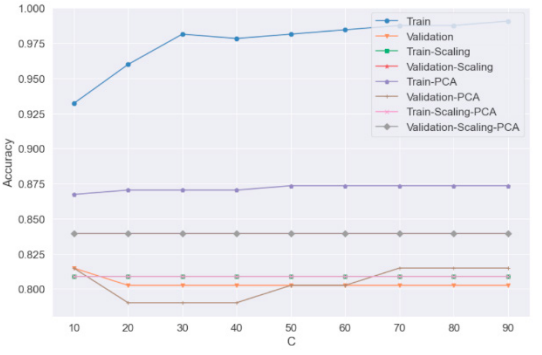
Algorithm	Types of the dataset	Tuning parameters	% Accuracy
KNN	Raw	Neighbors=4	83.95
	Standard Scaler	Neighbors=9	90.12
	PCA	Neighbors=6	85.19
	Standard Scaler + PCA	Neighbors=3	90.12
SVC	Raw	C=10	81.48
	Standard Scaler	C=10	83.95
	PCA	C=10	81.48
	Standard Scaler + PCA	C=10	83.95
LR	Raw	C=0.1	83.95
	Standard Scaler	C=0.79	91.36
	PCA	C=0.1	79.01
	Standard Scaler + PCA	C=0.1	90.12
LDA	Raw	Default	91.36
	Standard Scaler	Default	90.12
	PCA	Default	83.95
	Standard Scaler + PCA	Default	91.36
NB	Raw	Default	37.04
	Standard Scaler	Default	80.25
	PCA	Default	84.95
	Standard Scaler + PCA	Default	77.78
MLP	Raw	Neurons=200	83.95
	Standard Scaler	Neurons=650	93.83
	PCA	Neurons=250	83.95

	Standard Scaler + PCA	Neurons=50	88.89
RF	Raw	n_estimators=81, max_depth=91, min_samples_split=10, max_features=27	98.77
	Standard Scaler	n_estimators=91, max_depth=81, min_samples_split=10, max_features=27	98.78
	PCA	n_estimators=91, max_depth=21, min_samples_split=10, max_features=27	86.42
	Standard Scaler + PCA	n_estimators=61, max_depth=11, min_samples_split=10, max_features=27	91.36
DT	Raw	max_depth=71, min_samples_split=10, max_features=27	96.29
	Standard Scaler	max_depth=51, min_samples_split=10, max_features=27	97.53
	PCA	max_depth=81, min_samples_split=10, max_features=27	83.95
	Standard Scaler + PCA	max_depth=51, min_samples_split=20, max_features=27	90.12
KM	Raw	Clusters=16	83.02
	Standard Scaler	Clusters=3	92.90
	PCA	Clusters=15	82.72
	Standard Scaler + PCA	Clusters=3	92.90

Our results prove that the various algorithms work better by preprocessing and tuning parameters differently. Our findings show that KNN, SVC, LR, NB, MLP, RF and KM yield the highest accuracy after using Standard Scaler preprocessing. However, LDA and DT work better using the Raw dataset. Meanwhile, KNN, SVC, LR, LDA, MLP, RF, DT, KM yield the highest accuracy in the Standard Scaler+PCA dataset. Only NB produces the highest accuracy in PCA only. Overall, we can conclude that Standard Scaler and Standard Scaler+PCA help to obtain the best accuracy results. Parameter and preprocessing scaling can improve the accuracy of the algorithm.



A



B

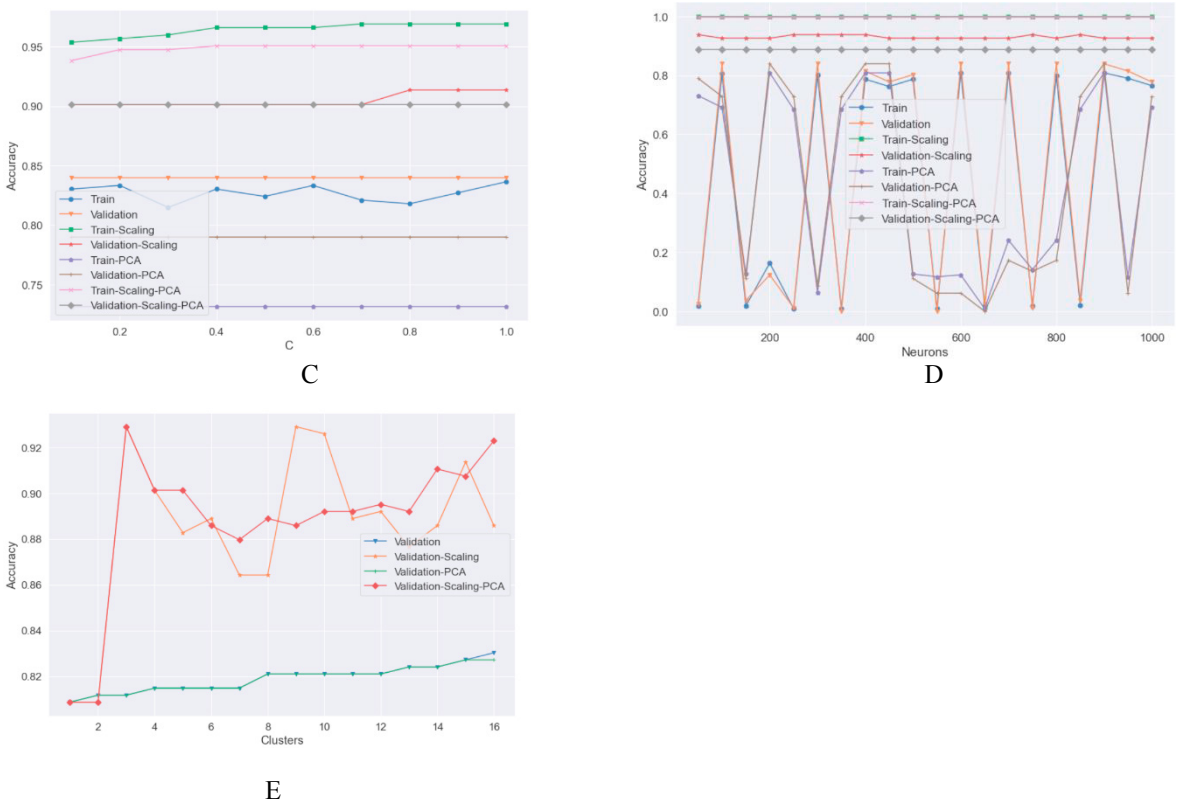


Fig. 3. Comparison of training and validation accuracy between parameters using all data set for (A) KNN, (B) SVC, (C) LR, (D) MLP and (E) KM.

Summarization of the result of the five best algorithms with the highest accuracy is listed in Table 4. Based on Table 4, the best algorithms are RF, DT, MLP, LDA and LR, where these algorithms achieved high accuracy near to 100%. All these algorithms used Standard Scaler datasets to retrieve the best accuracies.

Table 4. The best algorithms

Type of the datasets	Algorithm	% Accuracy
Standard Scaler	RF	98.78
	DT	97.53
	MLP	93.83
	LDA	91.36
	LR	91.36

5. Conclusion

Safe and clean groundwater is crucial in our daily lives, not only for our present lives but for future lives. Safe and clean groundwater is needed to sustain the public health, clean environments, and ecosystems. Therefore, early identification, robust prediction, and treatment are required to overcome the river pollution problems. Unsupervised and supervised ML algorithms are promising tools for the clustering and classifying uncertain and complex river pollution hotspots. Results show that obtained predictions with as high as 98.78% accuracies will allow

contributions to the discovery of accurate algorithms and river polluted hotspot areas for the early treatment and intervention.

Acknowledgements

This study was funded by the Malaysian Ministry of Higher Education (FRGS-RACER: RACER/1/2019/STG06/UNISZA//).

References

[1] National Geographic. (2021) "Surface Water." https://www.nationalgeographic.org/encyclopedia/surface-water/

[2] WWAP. (2015) "The United Nations World Water Development Report 2015: Water for a Sustainable World". UNESCO, Paris.

[3] UNESCO. (2021) "The global water quality challenge & SDGs." https://en.unesco.org/waterquality-iiwq/wq-challenge.

[4] Department of Environment. (2019) "Environmental Quality Report" https://enviro2.doe.gov.my/ekmc/wp-content/uploads/2020/09/EQR-20191.pdf.

[5] Elfikrie, N., Ho, Y. B., Juahir, H., and Tan, E. S. S. (2020). "Occurrence of pesticides in surface water, pesticides removal efficiency in drinking water treatment plant and potential health risk to consumers in Tengi River Basin, Malaysia". *Science of the Total Environment* **712**: 136540.

[6] Koki, I. B., Low, K. H., Zain, S. M., Juahir, H., Bayero, A. S., Azid, A., and Zali, M. A. (2020). "Spatial variability in surface water quality of lakes and ex-ming ponds in Malacca, Malaysia: the geochemical influence". *Desalination and Water Treatment* **197**: 319-327.

[7] Masthurah, A., Juahir, H., and Mohd Zanuri, N. B. (2021). Case study Malaysia: Spatial water quality assessment of Juru, "Kuantan and Johor River Basins using environmetric techniques". *Journal of Survey in Fisheries Sciences* **7**(**2**): 19-40.

[8] Wang, S., Peng, H., and Liang, S. (2022). "Prediction of estauarine water quality using interpretable machine learning approach". *Journal of Hydrology* **605**: 127320.

[9] Wang, R., Kim, J. H., and Li, M-H. (2021). "Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach". *Science and the Total Environment* **761**: 144057.

[10] Adusei, Y. Y., Quaye-Ballard, J., Adjaottor, A. A., and Mensah, A. A. (2021). "Spatial prediction and mapping of water quality of Owabi reservoir from satellite imageries and machine learning models". *The Egyptian Journal of Remote Sensing and Space Sciences* **24**: 825-833.

[11] Tabares-Soto R, Orozco-Arias S, Romero-Cano V, Segovia Bucheli V, Rodríguez-Sotelo JL, and Jiménez-Varón CF. (2020) "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data." *PeerJ Computer. Science* **6** (**e270**) : 1-22.

[12] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, and Telenti A. (2018) "A primer on deep learning in genomics." *Nature Genetics* **51** (**1**):12–18.

[13] Eraslan G, Avsec Ž, Gagneur J, and Theis FJ. (2019) "Deep learning: new computational modelling techniques for genomics." *Nature Reviews Genetics* **20** (**7**) : 389–403.

[14] Jabatan Pengairan dan Saliran Terengganu. (2021) "Water Pollution Statistics." http://jpsweb.terengganu.gov.my/

[15] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. (2011) "Scikit-learn: machine learning in python." *Journal of Machine Learning Research* **12** : 2825–2830.

[16] Hunter, J.D. (2007) "Matplotlib: A 2D graphics environment." *Computing in Science & Engineering* **9** (**3**) : 90–95.