An Online Voting System for Colleges and Universities

Article · May 2018

CITATION
1

READS
83,688

2 authors, including:

Idongesit Efaemiode Eteng
University of Calabar
32 PUBLICATIONS **16** CITATIONS

SEE PROFILE

UNIVERSITY OF THE WEST of SCOTLAND

School of Computing, Engineering & Physical Sciences

Computing and Information Systems Journal

Vol 22, No 2, 2018

Edited by Abel Usoro

www.uws.ac.uk

Computing and Information Systems is published normally three times per year, in February, May, and October, by the University of the West of Scotland.

From the next issue the editorial address is Dr Abel Usoro, School of Computing, University of Paisley PA1 2BE; tel (+44) 141 848 3959; fax (+44) 141 848 3542, e-mail: cis@uws.ac.uk or abel.usoro@uws.ac.uk.

Editorial Policy

Computing and Information Systems offers an opportunity for the development of novel approaches, and the reinterpretation and further development of traditional methodologies taking into account the rate of change in computing technology, and its usage and impact in organisations.

Computing and Information Systems welcomes articles and short communications in a range of disciplines:

- Organisational Information Systems
- Computational Intelligence
- E-Business
- Knowledge and Information Management
- Interactive and Strategic Systems
- Engineering
- E-Learning
- Cloud Computing
- Computing Science

The website for Computing and Information Systems is http://cis.uws.ac.uk

UWS UNIVERSITY OF THE WEST of SCOTLAND

Printed by Printing Services, University of the West of Scotland
University of the West of Scotland is a registered Scottish charity. Charity number SC002520

Text Classification Using Data Mining Techniques: A Review

Oluwakemi Christiana Abikoye[1], Samuel Oladeji Omokanye[2], Taye Oladele Aro[3]
Department of Computer Science,
University of Ilorin, Ilorin, Nigeria
[1]kemi_adeoye@yahoo.com, [2]oladejiomokanye@yahoo.com, [3]taiwo_aro@yahoo.com

ABSTRACT
Purpose: This paper gives an overview of data mining algorithms used for text classification and a review of works that have been performed on classifying texts.
Design/Methodology/Approach: Data mining algorithms used for text classification were discussed and researches done on applying such algorithms in classifying texts were considered with more emphasis on comparative studies.
Findings: No classifier can perform best in all situations as different datasets and conditions bring about different classification accuracies.
Practical Implications: In applying data mining algorithms for classifying text documents, it should be noted that the conditions of the data will affect classification accuracy; therefore such data should be well presented. Researchers may also need to try different algorithms and conditions to get a desired level of accuracy.
Originality/Value: A lot of work has been done in reviewing of data mining algorithms but this research has its specific emphasis on text and in addition to previous reviews, more recent journals were considered.
Keywords: *Data mining, Text mining, text classification, performance evaluation, classifier machine learning algorithm (MLA).*
Paper Type: *Research paper*

1 Introduction

There has been a wide availability of textual data which grows especially with increased size and accessibility to the World Wide Web, most of this data is not well structured in a way that makes mining of information from it automatically easy as a lot of these information are made available by different individuals and organizations through social media sites, news sites, review sites and various governmental and organizational agencies (Wang & Chiang, 2011). The relevance of Text mining is to be able to extract meaningful and useful information from the increasing available data so as to understand facts underlying such data and thus make good decisions for the betterment of the society (Dang & Ahmad, 2015). Text mining is the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents (Singh, 2016). All the extracted information is linked together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Its purpose is to get insights into large quantities of text data. The fundamental objective of text mining is to enable users to extract data from text based assets and manages the operations like retrieval, extraction, summarization, categorization (supervised) and clustering (unsupervised) (Dang & Ahmad, 2015).

Text classification is the task of categorizing a document under a predefined category. More formally, if di is a document of the entire set of documents D and {c1,c2, c3,...,cn} is the set of all the categories, then text classification assigns one category cj to a document di (Ikonomakis, Kotsiantis, & Tampakas, 2005). The documents depending upon their characteristics can be labeled for one class or for more than one class. If a document is assigned to only one class, it is called “single-label” and if the document is assigned to more than one class, it is called “multi-label” (Wang & Chiang, 2011).

A ‘‘single-label’’ text classification problem can be further categorized into a ‘‘binary class’’ problem if only one of the two classes is assigned to the document and this ‘‘single-label” text classification problem becomes a ‘‘multi-class’’ problem if only N mutually exclusive classes are assigned to the document. Text classification consists of document representation, feature transformation and/or feature selection, construction of a vector space model, application of data mining algorithm and finally an evaluation of the applied data mining algorithm (Jindal, 2015). Classification finds a model that separates classes or data concepts in order to predict

the classes of unknown objects. Take for instance a school will want to determine which of its final year student can be graduated, we have two categories, "graduate" and "spill" for the final year student data, the two categories can be represented by discrete values and the way it is ordered is irrelevant to the classification. Such is called supervised learning due to the fact that the classes in the training data have been labeled already.

A machine learning algorithm builds a classifier in two stages. (1) Training builds a classification model by analyzing training data that has class labels and (2) testing examines a classifier (using the test data) for accuracy and classify unknown objects into their respective classes (Sebastiani, 2002). A machine learning algorithms first builds the model to be used for classification by analyzing a training data which has class labels in it, then the classifier's model is evaluated by using a testing data. Such evaluation will be for accuracy in its ability to classify unknown data to its proper class, after which the classifier can then be deployed for real world use.

2 Text Classification Process

The process of classifying text involves identifying the data to be used for training the machine learning algorithm. Usually such data are in formats not suitable for mining or that will not give quality results. So such data is preprocessed through various preprocessing algorithms and sometimes, manual preprocessing needs to be done (Kotsiantis, 2007).

Preprocessing can also involve the removal of stop words, tokenization, lemmatization and stemming of words in the document, an expert need to have classified the training data into categories (for supervised learning) as it is such classification that the machine learning algorithm (MLA) will learn to form its classifier. The choice of which data part to be used for training or testing is also important; a large percentage of the dataset can be used for training while the remaining part can be left for testing. Cross validation can also be used, in which the data will be divided into equal sizes and a part will be left for testing while the other parts will be used for training. It is important not to use parts of the data used in training the MLA for testing as it will give a biased result.

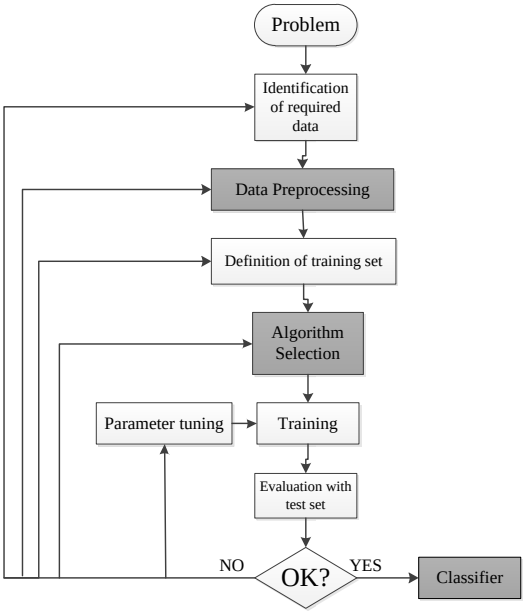

Fig 1: Text classification process (Kotsiantis, 2007).

The choice of algorithm to be used is also important; the nature and type of the dataset involved will have to be considered with respect to the characteristics of different machine learning algorithms as to make a choice. Different algorithms can be tested and the parameters tuned to get the best possible classifier. This shows that data mining is an experimental science (Witten, Frank, & Hall, 2011).

3 Machine Learning Algorithms used for Text Classification

There has been quite a number of MLA used for classifying texts. Some are based on logic such as decision trees and rule based classifiers, others are based on statistical principles such as Logistic regression, Naïve Bayes and Bayesian networks. There are also instance-based learning algorithms which learn when the actual classification is being performed, an example being the Nearest Neighbour algorithms. Others are Support Vector Machines and Artificial neural networks. Some of them are discussed in the rest of this section.

3.1 Naïve Bayes (NB)

Naïve Bayes is based on Baye's rule and it assumes independence naively (Witten et al., 2011); it multiplies probabilities of events assuming they are independent of each other. NB has been known to

work well with actual datasets and when combined with feature selectors which eliminate redundant and unimportant features, it works better (Witten et al., 2011), given a hypothesis H and evidence E,

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

3.2 K-Nearest Neighbour (KNN)

Nearest-neighbour classifiers learn by comparing a given test sample to the training samples similar to it (Jiang, Pang, Wu, & Kuang, 2012). KNN does not build a classification model as it is a lazy method (Jiang et al., 2012). It uses a distance function to determine which k members of the training set are closest to the unknown test instance (Witten et al., 2011). It predicts the unknown test instance using the majority class of the k members. KNN classifier works as follows (Jadhav & Channe, 2016).

Step 1: Initialize the value of K.
Step 2: Calculate the distance between the input sample and the training samples.
Step 3: Sort the distance.
Step 4: Take the top K-nearest neighbours.
Step 5: Apply simple majority.
Step 6: Predict the class label with more neighbours for input sample.

3.3 Support Vector Machines (SVM)

Support vector machines are algorithms which use linear models to implement nonlinear class boundaries by transforming the instance space using a nonlinear mapping into a new space, a linear model constructed in the new space can then represent a nonlinear decision boundary in the original space (Witten et al., 2011). SVM builds its method on the principle of VC dimension from statistical learning and Structural Risk Minimization (SRM) (Singla, Chambayil, Khosla, & Santosh, 2011). As described by Witten et al., (2011), SVMs are based on an algorithm that finds a special kind of linear model called the maximum-margin hyperplane. The instances that are closest to the maximum-margin hyperplane, the ones with the minimum distance are called support vectors.

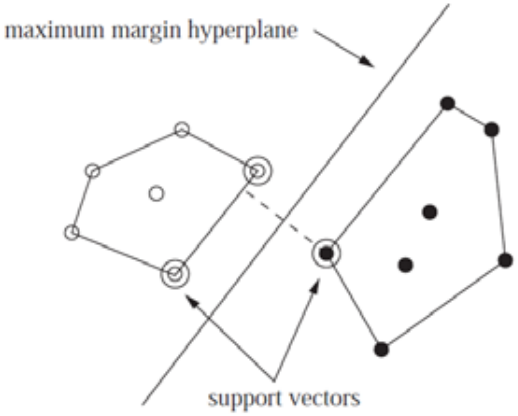

Fig 2: A maximum-margin hyperplane. (Witten et al., 2011)

3.4 Artificial Neural Networks (ANN)

Artificial neural networks are models which draw their inspiration from biological nervous systems which comprise of neural networks (Singla et al., 2011). ANN consists of highly interconnected network of an enormous number of neurons, an architecture inspired by the brain. As expatiated by Singla et al. (2011), Neural networks learn by examples; they are trained with known examples of the problem that knowledge is to be acquired from. When trained well, the network can be used effectively to solve similar problems of unknown instances.

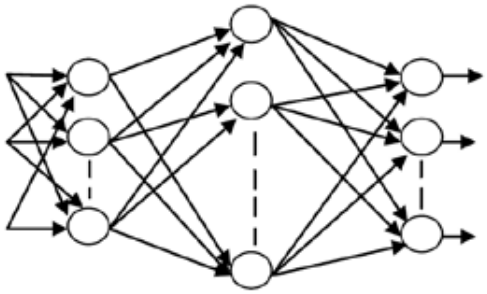

Fig 3: A three layered artificial neural network. (Shahare & Giri, 2015)

3.5 Decision Trees

A decision tree is a “divide and conquer” approach to the problem of learning from a set of independent instances (Witten et al., 2011). A tree is built using the data in the training set such that leaf nodes are a classification that applies to all instances that reach

the leaf or a set of classifications while for numeric predictions; the built tree is a probability distribution over all possible classifications. When an unknown instance is being tested for, the unknown instance is routed down the tree according to the attribute values tested in successive nodes and once a leaf is reached, it is classified as belonging to the same class as the leaf.

There are quite a number of decision tree based algorithms like Hunt's algorithm, CART, ID3, C4.5 (Patel & Rana, 2014).

3.6 Evaluating Algorithm performance

It is imperative to be able to evaluate an algorithm's performance so as to be able to determine its usefulness on the data it is being used to classify. Without assuming a particular application in mind, the ranking performance and binary classification performance would be informative for evaluating classifiers (Yang, 1999). There are some measures commonly used to evaluate the effectiveness of a classifier, they are as listed by Yang (2000) as precision, recall, F-measure, accuracy, and error. Others as described by Witten et al. (2011) are receiver operating characteristics (ROC) curve and Area under curve (AUC). Numeric predictions also have some other evaluators which are mean-squared error, root-mean squared error, mean absolute error, relative squared error, root relative squared error, relative absolute error and correlation coefficient. A detailed understanding of these numeric performance evaluators can be found in a book written by (Witten et al., 2011). Some of the commonly used evaluation measures are defined as follows:

Let TP = Documents correctly assigned to a category
FP = Documents incorrectly assigned to a category
TN = Documents correctly rejected from a category
FN = Documents incorrectly rejected from a category

$$Precision = \frac{\text{TP}}{TP + FP} \times 100$$

$$Recall = \frac{\text{TP}}{TP + FN} \times 100$$

$$F - measure = \frac{2 \times \text{Recall} \times \text{Precision}}{Recall + Precision}$$

$$Accuracy = TP + TN$$

$$Error = \frac{FP + FN}{TP + TN + FP + FN}$$

4 Related Work

Classification algorithms are being used by different researchers to perform classification tasks on datasets. A comparative study of some of the work is done, but as noted by Khan, Baharudin and Lee (2010), the reliability by algorithm comparison is majorly limited to experiments performed by the same author under carefully controlled conditions, because in comparisons of experiments by different authors, background conditions which are extraneous to the learning algorithms may affect the results in ways that are not reported.

Wang and Chiang (2011) proposed a method of finding multi-label categorization using SVM with membership function. Data mapping was performed to transform data from a high dimensional space to a low dimensional space with paired SVM output values, thus lowering the complexity of the computation. A pairwise comparison approach was applied to set the membership function in each predicted class to judge all possible classified classes. They compared their proposed model with several multi-label approaches which are Naïve Bayes, Multi-Label Mixture, Jaccard Kernel and Bp-MLL with their proposed method found to be better than these other ones in terms of overall performance indices.

Shugufta & Srinivasu (2017) applied SVM on reuter datasets using different combinations of training and test sets and discovered that the higher the number of training data the better the classification accuracy gotten.

Jain & Mishra (2016) proposed a model which combines NB with modified maximum entropy classifier. The two algorithms can be combined linearly by using its average, maximum or harmonic mean for classification of documents. They reported that the combination of the algorithms performs better than the individual algorithms.

Sallam & Hussein (2016) studied the effect of normalization and stemming on Arabic texts and from experimental results concluded that while normalization improves accuracy, stemming reduces accuracy with the base comparison being when neither normalization nor stemming was done.

Goudjil et al (2015) developed an active learning method for text classification which selects a batch of information samples to be manually labelled by an expert. Active learning using SVM was used to intelligently select which data were best labelled in

order to reduce the labelling effort without compromising accuracy. Results indicate that the proposed method significantly reduced the labelling effort and improved accuracy.

Jiang and Guasong (2012) presented a study which builds a classification model by combining constrained one pass clustering algorithm and KNN text categorization. The datasets used for their experiment are Reuters-21578, Fudan university text categorization corpus and Ling-Spam corpus. They used the clustering algorithm to compress and discover complex distribution of the training texts and the text documents are now classified based on the cluster vectors instead of original text samples by using KNN. This improved model is more effective and efficient than KNN and has significant performance and good generalization ability when compared with NB and SVM. It can also be incrementally updated which increases its applicability.

Shahare and Giri (2015) performed a comparative analysis of ANN and SVM classification for breast cancer detection. They used a dataset from the UCI repository composing 683 cytological instances of which 458 are benign and 241 are malignant cases of cancer. The performance evaluation was in terms of accuracy. Their results showed that SVM classifier has a better accuracy than ANN.

Jodas et al (2013) work also corroborated Shahare and Giri (2015) in their experimental results, as they compared SVM with ANN in recognition of steering angle for driving if mobile robots through paths in plantation. SVM gives an accuracy of 93% while ANN gives an accuracy of 90%.

Experimental results showing that SVM performs better than ANN in predictive accuracy was also supported by the work of Singla et al., (2011) when they compared the two algorithms performance in classifying eye events in electroencephalographic (EEG) signal.

Lu and Ling (2003) compared NB, Decision trees (C4.5 and C4.4) and SVM algorithms using both accuracy and AUC. Their experimental results showed that the predictive accuracy of the four algorithms are similar with no statistical difference between them while the average predictive AUC values of NB, C4.4 and SVM are similar, while C4.5 performed statistically lower. It is to be noted that their work was in the year 2003 so there have been significant improvements in those classifiers as recent literature suggests.

Yu and Xu (2008) performed a comparative study for content-based dynamic spam classification using NB, NN, SVM and relevance vector machine (RVM). They performed an empirical evaluation for them on the benchmark spam filtering corpora using different training size and extracted feature size. Their results posit that NN classifier is unsuitable for use alone as a spam rejection tool, SVM and RVM performed better than NB classifier, while SVM and RVM perform similarly, RVM is said to be more suitable for spam classification in terms of applications that require low complexity.

Zelaia, Alegria, Arregi and Sierra (2011) Presented a multi-classifier approach for multi-label document categorization problems, this was done by the use of a reduced vector representation obtained by singular value decomposition (SVD) for training and testing documents. Then a set of KNN classifiers were used for prediction. The KNN classifier uses the reduced database subsampled from the original database being trained from. A new approach based on Bayesian weighted voting was also introduced. Their results showed that constructing a multi-classifier jointly with the use of Bayesian voting to combine category label predictions brings about an improvement of results. Also the use of SVD dimensionality reduction technique greatly helped in reducing the vector representation of the document, as the original documents represented by 15,000 features in the Bag-of-Words form and by 11,000 in the Bag-of-Lemmas were simplified to 300 features thus saving time and space.

Hassan, Rafi and Shaik (2012) compared SVM and NB classifiers for text categorization with wikitology as knowledge enrichment. Using the 20 Newsgroup dataset, they evaluated the two algorithms using micro-average f-measure and macro-average f-measure. Compared to baseline results, SVM shows an improvement of +6.36% while NB shows an improvement of +28.78%, this shows that both classifiers are improved when information extracted from wikitology is integrated.

Colas and Brazdil (2006) used SVM, KNN and NB for their classification, the aim of their experiment was to examine the classifier learning abilities for an increasing number of documents in the training set using was the 20 newsgroup dataset. They performed binary classification tasks and came up with results indicating that although SVM has a good

overall performance, KNN with suitable preprocessing compares favourably and scale up well with the number of documents, NB also achieved a good performance.

Ashari, Paryudi, and Tjoa (2013) did a performance comparison between NB, Decision tree and KNN in searching alternative design in an energy simulation tool, the dataset used was gotten from raw building data available having 13 building parameters with each parameter having 4 possible values. The results obtained postulates that in terms of classification time, Decision tree is the fastest, followed by NB while KNN comes last. Decision tree fast classification time is attributed to the absence of calculations in the process of classification, while the slowness of KNN is attributed to the fact that the bigger the data, the larger the distance classification that must be performed, thus the time taken for classification is directly related to the number of data. Results also showed that NB outperforms the other two algorithms in performance.

While experiments are being performed on the actual classification tasks, others are being done on other aspects that can make classification easier, more effective and faster. Forman & Kirshenbaum (2008) proposed a method to speed up the feature extraction process that folds together Unicode conversion, forced lowercasing, word boundary detection and string hash computation, the model proposed results in classifiers with equivalent statistical performance to those built using string word features, but require far less computation and less memory.

Gonçalves, Souza and Gonçalves (2016) presented the first annotated dataset for the Brazilian savannah pollen types that can be used to train and test computer vision based automatic plain classifiers. The dataset contains 805 pollen images of 23 pollen types. They implemented a combination of three feature extractors and four machine learning algorithms.

Mohod and Dhote (2014) aimed at improving the feature selection method for text document classification in machine learning; the approach proposed for feature selection used the inverse document frequency divided by the document frequency. The experimental dataset was shown to perform well and thus is an alternative approach in selecting features in datasets.

5 Discussion

The use of machine learning algorithms for text classification is becoming more versatile as the available textual data is largely increasing and its potentials and practical use in marketing, diagnostics and prediction, mining the web amongst other applications. From the different journals reviewed, it is noted that different MLA algorithms performed differently with different datasets and conditions which implies that just trying to find an overall best classifier might not be effective. Research should be focused on finding conditions for which a classifier works better. This will help researchers in the process of making datasets available and having a better perspective of which algorithm to apply considering the situations at hand.

Evaluating how classifiers scale with data is also a viable area to research so as to have an idea of which algorithm is best to be used with the different data sizes.

A process of combining multiple algorithms together called ensemble learning to give a classification accuracy much better than the individual algorithms have been introduced over the last decade. It combines the strengths of the algorithms while complementing their weaknesses but has its own share of problems which as discussed by Kotsiantis (2007) includes increased storage which depends on the size of each classifier and that of the ensemble, increased computational complexity and comprehensibility.

Preprocessing techniques have also been found useful in helping classifiers predict better and faster. Studies on how to preprocess data are continuously a welcome idea.

6 Conclusion

Applications of machine learning algorithms on problems related to text are increasing so as to study patterns and correlations, and to extract useful information from such textual data. This overview discusses some popular algorithms used in text mining, and the metrics used in determining their suitability and performance. Previous works done on text mining especially as regards to the performance of algorithms in the classification of texts were also discussed. The comparisons show that a single algorithm cannot be best fit for all text classification

problems. Improving such algorithms to better classify is the focus of many researchers.

References

Ashari, A., Paryudi, I., & Tjoa, A. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *International Journal of Advanced Computer Science and Applications*, Vol. 4, No. 11, pp. 33–39. https://doi.org/10.14569/IJACSA.2013.041105

Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, Vol. 1, No. 1, pp. 4–20. https://doi.org/10.4304/jait.1.1.4-20

Colas, F., & Brazdil, P. (2006). Comparison of SVM and some older classification algorithms in text classification tasks. *IFIP International Federation for Information Processing*, Vol. 217, pp. 169–178. https://doi.org/10.1007/978-0-387-34747-9_18

Dang, S., & Ahmad, P. H. (2015). A Review of Text Mining Techniques Associated with Various Application Areas. *International Journal of Science and Research (IJSR)*, Vol. 4, No. 2, pp. 2461–2466. Retrieved from http://www.ijsr.net/archive/v4i2/SUB151800.pdf

Forman, G., & Kirshenbaum, E. (2008). Extremely Fast Text Feature Extraction for Classification and Indexing. In *Proceedings of the 17th ACM Conference on Information & Knowledge Management* (pp. 1221–1230). Napa, CA.

Gonçalves, A. B., Souza, J. S., Da Silva, G. G., Cereda, M. P., Pott, A., Naka, M. H., & Pistori, H. (2016). Feature extraction and machine learning for the classification of Brazilian Savannah pollen grains. *PLoS ONE*, Vol 11, No. 6, pp. 1–20. https://doi.org/10.1371/journal.pone.0157044

Goudjil, M., Koudil, M., Bedda, M., & Ghoggali, N. (2015). A Novel Active Learning Method Using SVM for Text classification. *International Journal of Automation and Computing*, pp. 1–9. https://doi.org/10.1007/s11633-015-0912-z

Hassan, S., Rafi, M., & Shaikh, M. S. (2011). Comparing SVM and Naive Bayes classifiers for text categorization with Wikitology as knowledge enrichment. In *Proceedings of the 14th IEEE International Multitopic Conference 2011, INMIC 2011* (pp. 31–34). https://doi.org/10.1109/INMIC.2011.6151495

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, Vol. 4, No. 8, pp. 966–974. Retrieved from http://www.math.upatras.gr/~esdlab/oldEsdlab/en/members/kotsiantis/Text Classification final journal.pdf%5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-23444448953&partnerID=40&md5=11a5f24b7ee05d580eccaf940e3499e4

Jadhav, S., & Channe, H. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR)*, Vol. 5, No. 1, pp. 1842–1845.

Jain, A., & Mishra, R. D. (2016). Text Categorization: By Combining Naive Bayes and Modified Maximum Entropy Classifiers. *International Journal of Advances in Electronics and Computer Science*, pp. 122–126.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, Vol. 39, No. 1, pp. 1503–1509. https://doi.org/10.1016/j.eswa.2011.08.040

Jindal, R. (2015). Techniques for text classification : Literature review and current trends. *Webology*, Vol. 12, No. 2, pp. 1–28.

Jodas, D. S., Marranghello, N., Pereira, A. S., & Guido, R. C. (2013). Comparing support vector machines and artificial neural networks in the recognition of steering angle for driving of mobile robots through paths in plantations. In *Procedia Computer Science* (Vol. 18, pp. 240–249). Elsevier B.V. https://doi.org/10.1016/j.procs.2013.05.187

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, Vol. 31, pp. 249–268. https://doi.org/10.1115/1.1559160

Lu, J., & Ling, C. X. (2003). Comparing Naive Bayes , Decision Trees , and SVM with AUC and Accuracy. In *IEEE International Conference on*

Data Mining (pp. 11–14).

Mohod, S. W., & Dhote, C. A. (2014). Feature Selection Technique for Text Document Classification : An Alternative Approach. *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 2, No. 9, pp. 2914–2917.

Patel, B. R., & Rana, K. K. (2014). A Survey on Decision Tree Algorithm For Classification. *International Journal of Engineering Development and Research*, Vol. 2, No. 1, pp. 1–5.

Sallam, R. M., & Hussein, M. (2016). Improving Arabic Text Categorization using Normalization and Stemming Techniques. *International Journal of Computer Applications*, Vol. 135, No. 2, pp. 38–43.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47. https://doi.org/10.1145/505282.505283

Shahare, P. D., & Giri, R. N. (2015). Comparative Analysis of Artificial Neural Network and Support Vector Machine Classification for Breast Cancer Detection. *International Research Journal of Engineering & Technology*, pp. 2114–2119.

Shugufta, F., & Srinivasu, B. (2017). Text Document Categorization using Support Vector Machine. *International Research Journal of Engineering and Technology*, Vol. 4, No. 2, pp. 141–147.

Singh, T. (2016). A Comprehensive Review of Text Mining. *International Journal of Computer Science and Information Technologies*, Vol. 7, No. 1, pp. 167–169.

Singla, R., Chambayil, B., Khosla, A., & Santosh, J. (2011). Comparison of SVM and ANN for classification of eye events in EEG. *Journal of Biomedical Science and Engineering*, Vol. 4(January), pp. 62–69. https://doi.org/10.4236/jbise.2011.41008

Wang, T., & Chiang, H. (2011). Solving multi-label text categorization problem using support vector machine approach with membership function. *Neurocomputing*, Vol. 74, No. 17, pp. 3682–3689. https://doi.org/10.1016/j.neucom.2011.07.001

Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical machine learning tools and techniques* (3rd ed.). San Francisco, CA,USA: Morgan Kaufmann Publishers Inc.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, Vol. 1, No. 1, pp. 69–90. https://doi.org/10.1023/A:1009982220290

Yu, B., & Xu, Z. (2008). A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge-Based Systems*, Vol. 21, No. 4, pp. 355–362. https://doi.org/10.1016/j.knosys.2008.01.001

Zelaia, A., Alegria, I., Arregi, O., & Sierra, B. (2011). A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. *Applied Soft Computing Journal*, Vol. 11, No. 8, pp. 4981–4990. https://doi.org/10.1016/j.asoc.2011.06.002

An Online Voting System for Colleges and Universities: A Case Study of National Association of Science Students (NASS), University of Calabar

Idongesit E Eteng
Department of Computer Science
University of Calabar, Nigeria
idongesitessien@yahoo.com

Ugochi D Ahunanya
Department of Computer Science
University of Calabar, Nigeria
ugochi_ahunanya@yahoo.com

Paul U. Umoren
Department of Computer Science
University of Calabar, Nigeria
umorenpaul@gmail.com

ABSTRACT

Purpose: This paper describes an online voting system that was designed to meet the electoral needs of universities and colleges.

Design/Methodology: The prototyping model was adopted as the methodology for designing the application. In designing the Online Voting System, Flowcharts, Use Case Diagrams and Data Flow Diagrams (DFD) were also employed.

Results: The system generated a more convenient voter and candidate registration interface, an efficient voting interface, vote storage and count plus immediate result compilation etc. Outputs from the application include a page showing a list of all the registered voters, a list of all qualified candidates, and the results of the total vote count for each candidate in the Faculty of Science. A functionality test was also carried out on the developed system where 20 registered students appraised the system by filling out an electronic questionnaire.

Originality: Several e-voting systems have been developed for varying uses. This system however was specifically developed for use in tertiary institutions and had security capabilities inbuilt into its design. This originality though peculiar to the adopted case study can be used for developing other kinds of applications. The system was also designed for faculty level voting but can be easily adapted for smaller or larger scenarios.

Practical Implications: It can be concluded that the Online Voting System incorporates all the features of a regular Voting system but offers an alternative method of conducting elections that is less stressful, easier and faster through the use of a network. It eliminates the moribund activities associated with the manual system and reduces drastically the duration of elections, thus, resulting in huge financial savings. It is thus recommended for use in any election if well adjusted.

Keywords*: Online Voting, Electronic Voting, Democracy, Voting System.*

Paper Type*: Research Paper*

1 Introduction

Elections are believed to be the key pillars of democracy and voting is one of the electoral processes that ensure the sustenance of democracy in any civil society. Online voting is an electronic way of choosing leaders via a web driven application (Amankona and Paatey, 2009). It provides a platform for simplifying the electoral process for all institutions that employ voting in decision-making. It is geared towards increasing the voting percentage in Universities and Colleges since it has been noted that with the old voting method (the Queue System or Manual System), the voter turnout has been a wanting case. In 2013, it was estimated that about 69 percent of students

stayed away from registering to vote. Of this figure, 60 percent claimed the reason they did not register was because they were not interested, 29 percent said they did not meet up the deadline while the remaining 11 percent claimed not to know where or how to go about the registration.

From the number of registered students, about 20 percent ended up not voting owing to many factors such as inconvenience, lack of faith in the process and for some a total loss of interest. (All statistics were compiled from data provided by the Electoral Commission of NASS, University of Calabar, Calabar, Nigeria). This statistics shows that the percentage of polling on the day of elections is definitely unsatisfactory and may not lead to electing the preferred candidate or the best person for the job.

The Online Voting System contains a database which is maintained and in which all the names of students with some basic information are stored. The individual votes are submitted in the database which can be queried to find out which of the aspirants for a given post has the highest number of votes. However, not just anybody can vote. For one to participate in the elections, he/she must have the requirements. For instance, he/she must be a registered student of one of the nine (9) departments of the Faculty of Science, University of Calabar. The registration should be done prior to the voting date to enable data update in the database.

The Online Voting System calculates the results faster, reduces time spent making long queues at the polling stations during voting, reduces human efforts, and also enables voters to cast their votes from any part of the globe. Occurrences of vote miscounts was drastically reduced since at the backend of this system resides a well-developed database. It has several security requirements like access control as well as user authentication incorporated into its design structure, making it have a higher measure of security, reliability, and resilience. It also provides for user-friendly graphical interfaces and tools which make voting easy and enjoyable.

1.1 Background of Study

The University of Calabar - also known as UNICAL - is a university situated in Calabar, Cross River State, Southeastern Nigeria, and is one of Nigeria's second generation universities. The University of Calabar is in the forefront of information technology drive demonstrated in many ways such as provision of access to Internet services for staff and students in the school area and amongst other things, launching an online-based transcript system for transcript application processing. The University is also one of the foremost Nigerian Universities to automate students' registration processes through the College Portal.

The National Association of Science Students (NASS), University of Calabar, is a student body that seeks the welfare of the students in the Faculty of Sciences (one of the about twelve (12) faculties of the University). The Faculty of Sciences comprises of nine (9) departments. In NASS, general elections are carried out every session where the President and nine (9) other executive members are elected by the students of the faculty and conducted by the NASS Electoral Commission (NASS Eleco).

Most Electoral bodies (NASS Eleco not an exception; right from their inception to date and even with latest advancements in technology) still use a primitive paper based method during voting. This system is characterized by manual form filling to choose leaders and transfer of the information from manual data capture forms to computerized datasheets. This has led to an excessive number of mistakes making their way into the final vote counts hence leading to confusion at the time of announcing the results.

The main advantage of paper-based systems is that ballot papers are easily human auditable. The disadvantages outweigh the advantages. For instance, the need to print ballot papers is a slow, expensive, inflexible, environmentally hostile process. Also, visual impairments, or literacy limitations plus last minute changes to the voters' register are difficult to accommodate amongst others.

Over the last few years, a number of election observers have suggested that electoral organizations introduce electronic/online voting (Sabo, Siti and Rozita, 2015). A general observation is that as more business is done using electronic media, it should not be difficult to carry out voting using electronic equipment rather than turning up at the polling place on the voting day to use paper and pen. Evidently, the phenomenal use of the Internet as a vehicle for improving

communication, access to information and electronic commerce has led to the claim that the Internet could be used as either a replacement to attendance voting or as an additional voting option (Suleiman, and Gwani, 2015).

1.2 Purpose of Study

This paper describes an online voting system that was designed to meet the electoral needs of universities and colleges, and also tackles the inherent problems of the present manual voting system. This current system—manual voting—is characterized by absenteeism, inconvenience, long queues, stress, a lot of paper work, error-prone human effort involved in vote computation, omissions, delays and other election irregularities which plague the system and defeat the whole aim of voting (Kohno et al , 2004).

In view of the rapid development of computer technology in virtually all fields of operations and its use in relation to information management, it has become pertinent to look into the development of an Online Voting System that can achieve the following:

- Conduct free and fair elections.
- Safeguard data and information in the system.
- Reduce workload in the process of conducting elections.
- Keep accurate record of votes.
- Reduce time wasted in announcing election results.
- Eliminate disenfranchising electorates.

The objectives of the proposed Online Voting System for Colleges and Universities is to use computer technology or information technology to simplify the electoral process and to

- Review the existing/current voting process or approach in colleges/universities;
- Design an automated voting system that should be able to handle extremely large volumes of data;
- Implement an automated/online voting system that should support multi-user environment;
- Validate the system to ensure that only legitimate voters are allowed to vote.

2 Literature Review

2.1 Introduction

The Online Voting system is made for the students to be able to vote for their representatives from any part of the globe. It is done on the Internet and as such can also be called the Internet Voting (Kuye et al, 2013). It seeks (or should seek) to accurately reflect the voters' preferences.

Online voting systems are appealing for several reasons but mostly because people are generally getting more acquainted with using computers to do all sorts of things, namely sensitive operations such as shopping and home banking, and it allows people to vote at their convenience, helping to reduce the rate of absenteeism which ranks as the highest malady plaguing the electoral process world over.

2.2 The Concept of Electronic Voting

The term "Electronic voting" has been used for a large variety of systems, ranging from hand-held infrared devices and kiosk systems with touch screen machines used in polling stations, to remote voting via the Internet. This paper is focused on remote voting via the Internet solely, using a computer system. E-Voting is the preferred platform for future elections in the developed and developing nations of the world (Morley and Parker, 2007). It is a system that has modernized the electoral processes and electorates are able to cast their votes through an electronic device as against the traditional manual system (Grossman, 2004).

Electronic voting (also known as E-voting, Online Voting, I-Voting, and Internet Voting) is a term encompassing several different types of voting, embracing both electronic means of casting a vote and electronic means of counting votes. Electronic voting technology can include punched cards, optical scan voting systems and specialized voting kiosks (including self-contained direct-recording electronic voting systems, or DRE). It can also involve transmission of ballots and votes via telephones,

private computer networks, or the Internet (Idike, 2014)

Generally, three main classes of E-Voting can be identified:

- **Polling station E-Voting/Poll-site Internet voting:** where voters cast their votes electronically on an electronic machine within the polling booth and voting is physically supervised by representatives of the electoral authorities.
- **Kiosk E-Voting**: where voters cast their votes at pre-selected stations through ATM-Iike terminals; Voting machines would be located away from traditional polling places, in such convenient locations as malls, libraries, or schools. The voting platforms would still be under the control of election officials, and the physical environment could be modified as needed and monitored (e.g., by election officials, volunteers, or even cameras) to address security and privacy concerns, and prevent coercion or other forms of intervention.
- **Remote e-Voting**: where voters cast their votes anywhere and anytime there is Internet access; as well as voting through mobile devices. Here, voting is performed within the voter's sole influence, and is not physically supervised by representatives of governmental authorities (also called Internet-voting or Online voting). It seeks to maximize the convenience and access of the voters by enabling them to cast ballots from virtually any location that is Internet accessible (Sabo, Siti and Rozita, 2015).

2.3 Types and Variation of Voting

Paper-based voting: this is the most common type of voting system where a voter gets a blank ballot and uses a pen or a marker to indicate which candidate he/she wants to vote for. It is both time and labor consuming, but it is easy to manufacture paper ballots and the ballots can be retained for verification. This paper ballot system was first adopted in the Australian state of Victoria in 1856 and the paper ballot system thereafter became known as the "Australian ballot." (Vermont, 2011)

Mechanical Lever Voting Machine: A Lever machine is a peculiar equipment, and each lever is assigned to a corresponding candidate. The voter pulls the lever to poll for his favorite candidate. This kind of voting machine can count up the ballots automatically. Since its interface is not user-friendly enough, giving some training to voters is necessary. (Vermont, 2011)

Direct Recording Electronic Voting Machine: This type—which is abbreviated to DRE—integrates with keyboard, touch screen, or buttons for the voter to press in order to poll/vote. They are an electronic implementation of the old mechanical lever systems. As with the lever machines, there is no ballot; the possible choices are visible to the voter on the front of the machine. The voter directly enters choices into electronic storage with the use of a touch-screen, push buttons, or similar devices. An alphabetic keyboard is often provided with the entry device to allow for the possibility of write-in votes. The voter's choices are stored in these machines via a memory cartridge, diskette or smart card and added to the choices of all other voters. DREs can come with or without a paper trail (VVPAT, or voter-verified paper audit trail). VVPATs are intended to provide physical evidence of the votes cast. (Vermont, 2011)

Punch card: The voter uses metallic hole-punch to punch a hole on the blank ballot. It can count votes automatically, but if the voter’s perforation is incomplete, the result is probably determined wrongfully.

OMR systems which are based on scanners that can recognize the voters’ choice on special machine-readable ballot papers. OMR systems can be either central count systems (where ballot papers are scanned and counted in special counting centers) or precinct count optical scanning (PCOS) systems. (Gordon, 1998)

Internet voting system or Online Voting System where votes are transferred via the Internet to a central counting server. Votes can be cast either from public computers or from voting kiosks in polling stations or—more commonly—from any

Internet-connected computer accessible to a voter. (Gordon, 1998; Thomas and Markus, 2005)

2.4 Characteristics of a Good Voting System

Voting systems must be transparent and comprehensible enough that voters and candidates can readily accept the results (Kohno et al., 2004). This means that the veracity of a voting system is necessary for the acceptance of the results of that election.

- Shamos (2004) gives a comprehensive assessment and states that for a voting system to be considered transparent and comprehensible, some important criteria must be met, otherwise it may lead to indecisive or inaccurate election resultsFirst of all, the anonymity of a voter's ballot must be preserved, in order to ensure that the voter is safe when voting against a candidate, and also to guarantee that voters have no evidence that proves which particular candidates received their votes.
- Secondly, the voting system must be tamper-proof in order to prevent a wide range of attacks both by voters and by insiders (poll officials).
- Thirdly, it should be user friendly. This means that it should be easily comprehensible and usable by the entire voting population.

2.5 Historical Development of Voting

Countries progressed considerably from the raising of hands to the use of machines in the voting process. The early paper ballots of the 1800s had no standardization whatsoever. Voters could add names to the ballots and because there was no method to verify the identity of the voter, voters could even vote at multiple locations. The United State recognized a need for a standardized ballot in 1880, yet they did not implement Australia's White Paper Ballots until 1888. (Saltman, 2006). The type of machines in the voting process has evolved from a mechanical type to an electronic type, falling in line with technology advances. It is interesting to note that there was no paper audit trail available to recount votes, but by this time, voters did have to verify their identity and sign their names in a book. Nevertheless, a lot has changed over the years.

In their paper, Sanjay et al (2011), give a thorough analysis of electronic voting in various countries. Some countries whose voting processes were described include: Brazil, India, Belgium, Australia, Italy, Argentina, the United Kingdom, Costa Rica, Panama and Spain. The factors considered for comparing these nations included:

i) whether a country's system uses a paper audit trail.
ii) whether the system permits an anonymous, blank or spoiled ballot.
iii) whether the software is open source or proprietary.

In the United Kingdom for example various technological improvements to voting or vote counting have been tested and applied in different forms of elections. Advancements include the use of technologies such as touch-screen voting machines while others tested techniques for voting remotely. Some jurisdictions even permitted voters to cast their ballots using electronic methods, such as interactive voice response (IVR) technology, Personal Computer based systems and handheld mobile devices via short message service (SMS). Some of these jurisdictions allowed voters to cast ballots from PCs or kiosks in public places such as shopping centres. Our work was however designed to meet the voting needs of higher institutions and is similar to the work done by Quist et al (2016)

Even though the electronic system described in this work was developed for academic institutions, it can be adapted for other forms of voting exercises if well adjusted.

3 Methodology

A six step prototyping model was used for the development of this application. The phases of the development model included:

Requirements Gathering
Quick design
Prototype Building
Product Engineering
Prototype Refinement
Customer Evaluation

Requirements were gathered from the Faculty members, student association and the constitution

of the association. The rest of the steps are detailed below.

3.1 The Framework of the Online Voting System

The diagram in figure 1 shows the framework of Online Voting System. In this online process of voting, a student (user) fills the registration form to register.

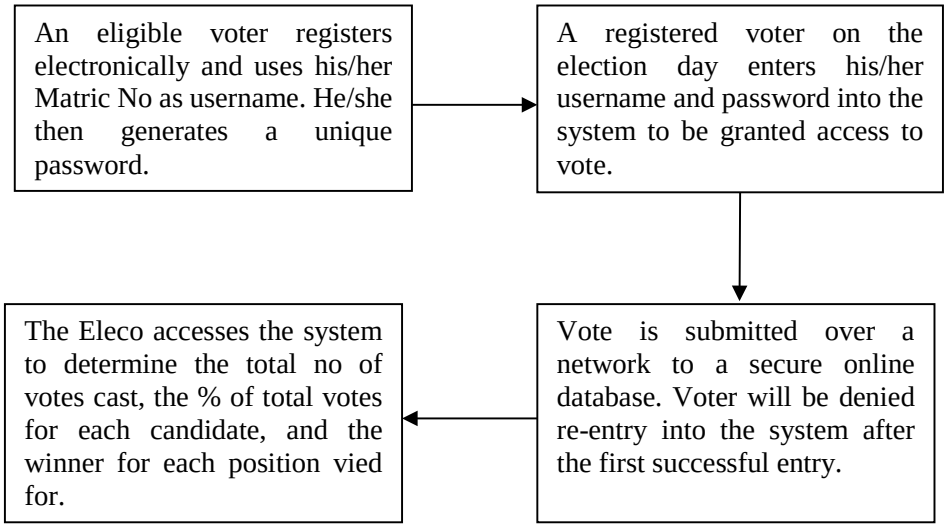

Figure 1: The framework of Online Voting System

The details filled in the form are checked with valid details obtained from the departments and preloaded in the database. If a match is found then username will be the matriculation number while the password will be created by the user to execute his/her franchise. Otherwise, the transaction is discarded. The users in this project include voters, candidates, ELECO and Administrators handling the system. By using information technology, it is believed that the proposed system will enable votes to be cast and counted with higher convenience, efficiency, improved performance and security. Hopefully, it will thus simplify the electoral process while easing the work pressure of manually maintaining details and consequently reduce the mistake rate of ballot election.

.

3.2 Quick Design and Prototype Building

3.2.1 User Requirements for the Proposed System

The Online Voting System should:

- Be able to display all registered voters in the database to the Admin and Eleco
- Have a user-friendly interface and user guides understandable by people of average computer skills.
- Be robust enough so that users do not corrupt it in the event of voting.
- Be able to handle multiple users at the same time and with the same efficiency
- Be scalable (for future expansion)

3.2.2 Security Requirements

The security requirements of the system are:

- An individual not registered to vote must not be able to cast a ballot.
- A voter must not be able to vote more than once.
- The privacy of the vote has to be guaranteed during the casting, transfer, reception, collection, and tabulation of votes.
- No voter should be able to prove that they voted in a certain way.
- None of the participants involved in the voting process (organizers, election officials, trusted third parties, voters, etc.)

should be able to link a vote to an identifiable voter.

- Each vote is recorded precisely as the voter intended.
- Each voter is ensured a "clean slate" of the system to ensure equality and confidence.
- The outcome of the voting process must correspond to the votes cast.
- It should be infeasible to exclude a valid vote from the tabulation, and to validate a non-valid one.
- System operations are logged and audited.
- The system cannot be re-configured during operation.
- Access to voted ballots is prohibited until after the close of polls.
- Additional ballots cannot be cast once the polling has ended.
- The system must be open to independent inspection and auditing.
- The system should be protected against accidental and malicious denial of service attacks.

3.2.3 Input Requirement

The input design depends on the type of output required. It involves data collection methods and validation which is done online or offline. It is structured and interactive. The major input requirements of the online voting system are listed below:

Cleared Student Information: This holds information about the students that have been cleared in their various departments uploaded into the database.

Table 1: Cleared Student Information Input Design

Id	Matric. No	Name	Department	Course	Level

ELECO Information: This is basically information about the students that make up the Electoral council in charge of conducting the election.

Table 2: Eleco information input design

Id	Surname	Other Names	Username	Password	Matric. No	Dept	Passport	Position

Other input design tables include: Candidates Information Input Design, Submission of votes Information Input Design, Output Requirement Input Design, Qualified candidates to be voted for Input Design, Registered students Input Design, Result of election Input Design.

3.3 Design Tools

In this system design, some important design tools required to aid the development of the proposed system include use case diagrams, dataflow diagrams, flow charts and database design.

3.3.1 Use Case Diagram

Below is the use case diagram that describes the system.

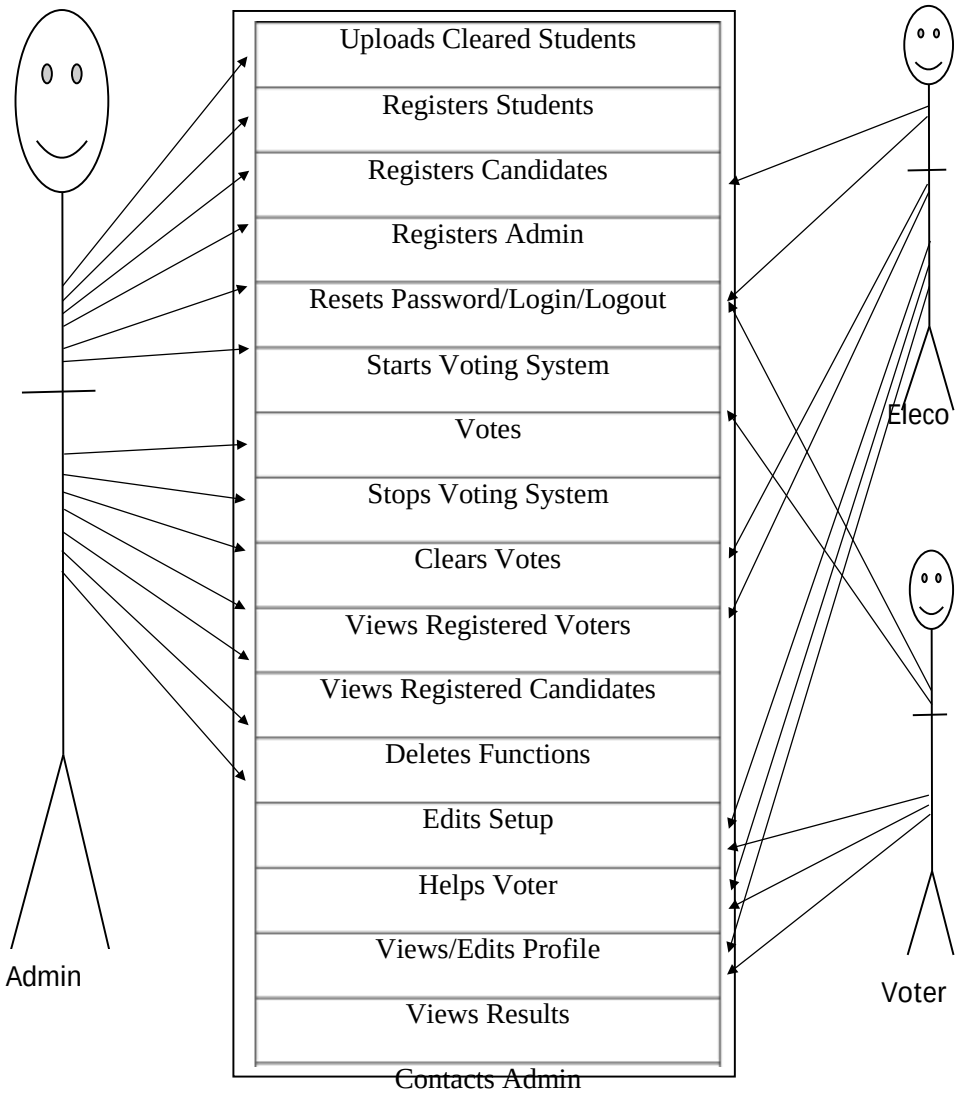

Figure 2: Use Case Diagram

3.3.2 Online Voting System Dataflow Diagrams

Below are the data flow diagrams.
They describe the flow of data

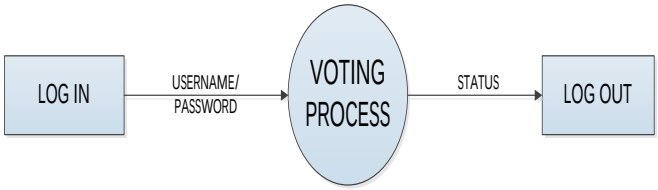

Figure 3a: Level 0 Dataflow Diagram for Online Voting

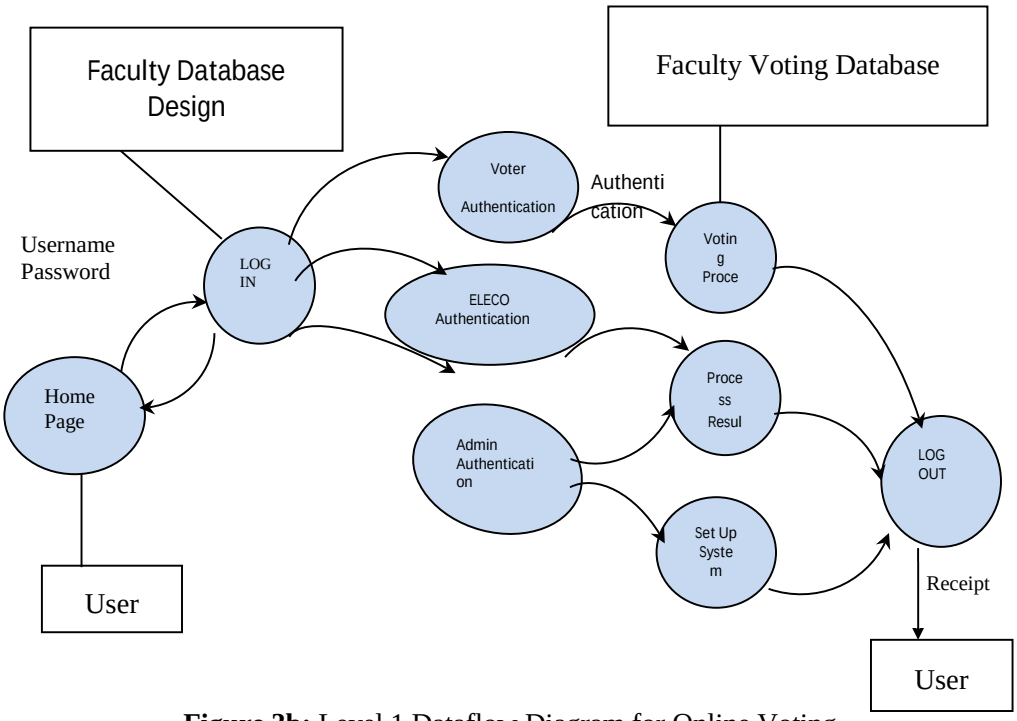

Figure 3b: Level 1 Dataflow Diagram for Online Voting

3.3.3 Database Design

The Online Voting System uses a database called Nass_e_Voting which comprises of seven (7) tables as illustrated below:

Admin table holds records of the administrators who control the system.

Table 3: Admin table

Field	Type	Null	Default	Primary Key	Auto-increment
adminId	int(11)	No		Yes	Yes
Username	varchar(16)	No			
Gender	varchar(10)	No			
Secret code	varchar(50)	No			

Table 3 illustrates the table structure for Admin table with its associated fields.

Eleco table holds records of the electoral officials who are in charge of coordinating the voting process. The fields are illustrated in the diagram below:

Table 4: Eleco Table

Field	Type	Null	Default	Primary Key
Id	int(10)	No		Yes
Surname	varchar(20)	No		
other_name	varchar(30)	No		
Username	varchar(30)	No		
Password	varchar(50)	No		
matric_no	varchar(12)	No		
Department	varchar(30)	No		
Passport	varchar(30)	No		
Position	varchar(20)	No		

Table 4 shows the structure for Eleco with its clearly defined fields. Other tables include candidate table, cleared students table, students table, user table, and vote tracker table.

3.3.3 *VOTER's Flowchart Diagram*

The main flow chart diagram used for the application is shown in Figure 4. Other flowcharts modelled are ADMIN'S Flowchart and ELECO'S flowchart.

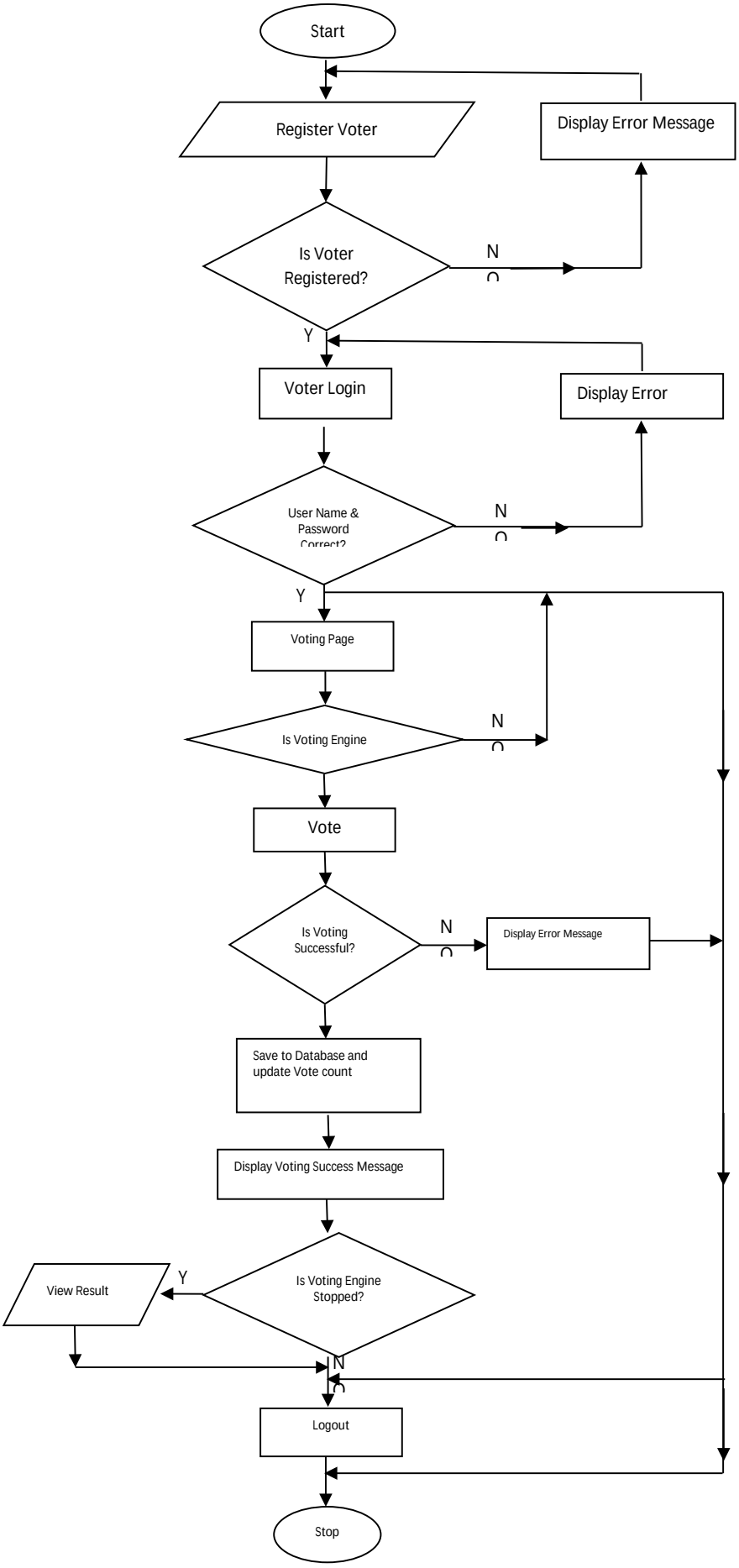

Figure 4: Application Flowchart

4. RESULTS

4.1 Implementation of the Proposed System

System implementation is the practice of creating or modifying a system to create a new or replace an existing business process. It consists of converting hardware and files to the new system and also of training the user on how to handle the system.

The system was developed as an interactive mechanism between the user at the interface and the database using the web-browser. This tool enables a user through a web browser to interact with the MYSQL database to enter, edit, view and retrieve such data as per the privileges granted. HTML forms offer the best layout to enter data, change and view the database. These forms were also kept as short and simple as possible for easy public awareness on the use of the tool. The interfaces and forms created include the following:

4.1.1 The Home Page/User Login Page

On visiting the Online Voting System site, this is the first page the user interacts with. It can also be called the Index page. It serves as the starting point to prospective users (Eleco and Voter). The individual is required to register and/or provide a username and password for authentication into the system.

Figure 5 shows a snapshot of the User Login Page.

Figure 5: User Login Page

4.1.2 The Candidate Registration form Page

This form is strictly reserved for the system administrator. He/she is the only one with the privilege(s) to access and use this form to register prospective candidates.

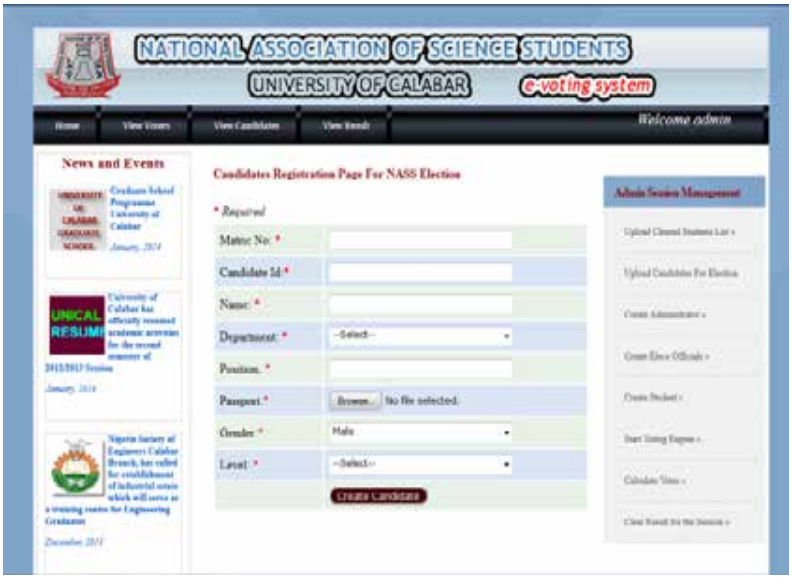

Figure 6: Candidate Registration Form Page

The Figure 6 illustrates the Candidate Registration Form Page

4.1.3 Login Page

Each user once logged in, is a legitimate user of the system and therefore given the privilege to perform functions such as edit profile, contact admin, start voting and view result.

Figure 7: Login Page

The Figure 7 shows the login page of the Online Voting System

4.1.4 *Voting Page*

This is the voting page where the actual voting is done. Only a legitimate voter can cast his/her votes. Once a vote has been cast by a user, he/she cannot go back to vote for the same position.

Figure 8: Voting Page

The Figure 8 shows a snap shot of the Voting Page of the Online Voting System

4.1.5 Admin Login Page

This page captured in Figure 10, authenticates an administrator into the site.

Figure 9: Admin Login Page

4.1.6 Admin Home Page

This page captured in fig 4.6 welcomes the administrator after authentication into the site. It contains the necessary functions of the administrator such as uploading cleared students, registering candidates.

Figure 10: Admin Login Page

Figure 10 displays a snapshot of the Admin Login Page.

4.1.7 The Result Page

After voting, a voter of Eleco is allowed to check the results by visiting the results page. This can only happen when the voting period has elapsed and the voting engine is stopped.

Figure 11: Results page

A snapshot of the Result Page is illustrated in Figure 11.

4.1.8 The Voting Error Page

The page captured in Figure 9 shows when a voter has successfully submitted a vote for a particular position and wants to vote for the same position again.

Figure 12: Voting Error page

A snapshot of the Voting Error Page is illustrated in figure 12.

4.2 User Manual

Double click on the Internet Explorer Icon located on the desktop and launch the web browser or any web browser of choice. Type the following address in the address bar: www.unical.nass-e-voting.edu.ng and click on "Enter" key.

Click on "not registered" link if you are a new user or proceed to step 4.

Fill the student registration form and click "register". If you have been cleared by your department and your name has been successfully uploaded to the database, you will get a confirmation message saying that your registration is successful but if not, you'll get an error message.

Login with your user name and password and then choose the category (voter or eleco) and click submit

As a voter, you are now logged in and can edit your profile, contact admin or vote when the voting engine has been turned on.

To edit your profile, click on the edit profile link and fill the form correctly with your correct details. Click on the "save changes" button to save the changes made.

To contact Admin, click on the "contact admin" link and submit you comment or complaint or get the numbers to call.

To vote, Click on "start voting" and select the link to vote for the candidate of your choice. It will take you to the candidate's page for each position. A confirmation message will be received each time a vote is done. Note that once you have submitted the vote to a particular position, you cannot go back to make changes.

Click on the particular candidate of choice and click on the "click here to submit vote button".

Click on the logout button to log out when through.

4.3 Testing

Testing is the process of running a system with the intention of finding errors or checking if it is working well by meeting the parameters established for it. The Online Voting System was designed to run on a web browser thus a WAMP server (which is basically built to test websites to see if the expected functionalities are met before finally uploading it online) was installed to test the system and was confirmed to be logically and functionally correct and working as shown in the snapshots illustrated in Section 4.1.

4.4 System Evaluation

The system was hosted and evaluated by 20 registered students. The basic test was a functionality test. Students were made to use the system and to fill some questionnaires that contained questions used as test cases. A few of the responses were analyzed using SPSS.

Excerpts of the results are given in tables 1 to 7.

Table 5: Admin Evaluation of Votes
(Able to add ballot boxes after pooling places closed as an Admin)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Passed	15	75.0	75.0	75.0
	No response	5	25.0	25.0	100.0
	Total	20	100.0	100.0	

Table 6: Admin Security Check
(Observed any form of attack during system usage as admin.)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	passed	20	100.0	100.0	100.0

Administrator has the privilege to upload an Excel formatted document containing eligible voters.

Table 7: Admin Upload Check

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	passed	20	100.0	100.0	100.0

4.4.1 Security Features Built-in in the System

The following security features were built into the system:

a) An individual not registered to vote is not able to cast a ballot.
b) A voter is not able to vote more than once.
c) Password authentication: This is another good security feature for checking all unauthorized attempts.
d) Authentication of session and session timeout has been provided to prevent session hijacking.
e) Password has been hashed (one way encryption) as shown in Figure 5.
f) System ensures that voter's password is 8 characters or more.
g) System also ensures that the username entered by the voter during registration is 8 characters or more.
h) System does not at any time show voter's password in plaintext.

5. Conclusion and Recommendations

5.1 Conclusion

This paper has come as a platform to propose an online voting system that will place our democracy on a path of success. In a nutshell, this research seeks to increase the efficiency of voting process and the image of the Electoral Commissions in charge of elections.

This Online Voting System is simple and indigenous. It has an authentication feature that will manage the Voter's information by which voter can login and use his/her voting rights. There is a database in which all the names of voters with complete information is stored and it provides the tools for maintaining voter's vote to every candidate and also computes the total number of votes of every candidate. Online voting offers speed and convenience to the voter and considerable ease to election administrators as they can get election results out more quickly than conventional methods of manual voting since vote counting is just a matter of querying the database.

The success factors include:

- Faster electoral process.
- A better platform for the disabled as well as the outright elimination of multiple voting.
- Finally, the integrated system would avail the electorates the opportunity of casting their votes using the most convenient means.
- The adoption of this system is likely to increase the level of participation in the polity because of the ease of voting and its tendency to eliminate electoral fraud.

5.2 Recommendations

The recommendations that came out of the research are presented in this section for consideration by systems developers.

5.2.1 Get key stakeholders to buy in.

As introducing Online voting is a trade-off of advantages and disadvantages, make sure that there is wide agreement among stakeholders, that this technology is overall advantageous.

Be aware that significant opponents of the system can and will come up with objections, and the weaknesses of the system can create distrust in the system and potentially in the entire electoral process. Even in the absence of genuine opposition to Online voting, the system can become disputed for purely political reasons.

5.2.2 Provide for transparent auditing and certification.

Online Voting Systems should be certified by an independent agency and audits should be conducted throughout the process to allow independent confirmation of the results produced.

Certification and audits are important confidence-building measures and should be transparent, allowing stakeholder's access to related procedures and documentation.

5.2.3 Plan for training, professional development, and civic and voter education.

Well-informed voters will not only find it easier to use e-voting on election day; they will also find it easier to trust a new system if they understand why it is being introduced, what benefits it brings and the various security measures that are built to support the integrity of the election.

5.2.4 Consider sustainability issues and plan for the future, not only for today.

The cost of introducing an online voting can already be very high, but to remain secure and trustworthy online voting systems need continuous reviews, upgrades and replacement as well as adjustments to new requirements. When considering the costs of e-voting it is important to consider the total cost of maintenance over time rather than the one-time purchase costs.

5.2.5 Infrastructural Review.

The standards of ICT infrastructure in the School and the country at large should be reviewed and developed.

5.2.6 Use of Biometrics

There should be use of biometric capturing devices, which will serve as a means of voter's authentication. There should also be adequate and proper public enlightenment before the system is fully implemented.

5.2.7 The Online Voting System can be used for different elections.

In this project we handled election for just the Faculty of Science, University of Calabar, but this same system can be used in future for conducting different elections like Students Union Government (SUG), Company elections and even national elections. The only requirement is that we need to create the whole voters database.

5.3 Further Study

An integrated voting system that incorporates an Electronic Voting Machine (EVM), Internet Voting (i-Voting) and Mobile Voting (m-Voting) is proposed for enhanced participatory democracy.

The use of Biometrics (Finger print/cornea detection), CAPTCHA and immediate password change could also be incorporated into the system in future to enhance the security strength though there is a problem of it decreasing the scope of the platform because these systems need some electronic components to implement. So, it will avoid the users' privilege to cast the votes at their fingertips. But it can guarantee that fake voting will be impossible.

SMS Update: In future SMS query could also be added. This will enable the sending of registration details or result updates to mobile phones or hand-held devices.

References

Amankona, E. and Paatey, E. (2009). Online Voting Systems. Graduation Project, Wisconsin International University College, Ghana.

Gordon, T. (1998). On Voting. *Kyklocls International Review for Social Sciences*, Vol 52, No 1, pp 126–128.

Grossman, W. M. (2004). Ballot Breakdown. *Scientific American*, Vol 290, No 2, pp 16-18.

Idike A. N. (2014). Democracy and the Electoral Process in Nigeria: Problems and Prospects of the E-Voting Option, *Asian Journal of Humanities and Social Sciences* (AJHSS), Vol. 2, Issue 2, pp 133-141. Available online at www.ajhss.org [Accessed: 10 Apr 2015].

Kohno T., Stubblefield, A., Rubin, A. D., Wallach D. S. (2004). Analysis of an Electronic Voting System, *Proc. 2004 IEEE Symp. Security and Privacy*, IEEE Press, pp 27–40.

Kuye C.O., Coker J.O., Ogundeinde I.A. and Coker, C.A. (2013). "Design and Analysis of Electronic Voting System in Nigeria", International Archive of Applied Sciences and Technology, Vol 4, No 2, pp 15-20. Available online at www.soeagra.com/iaast/iaast.htm [Accessed: 19 Jan 2015].

Quist, S. C., Amagate. L. K., Dickson, D. Conceptual design of e-voting System for Academic Institution (2016). *International Journal for Innovation Education and Research.*Vol. 4, No 12. pp 96-108.

Saltman, R. (2006). Independent Verification: Essential Action to Assure Integrity in the Voting Process. *National Institute of Standards and Technology (NIST) Report SB134106W070*.

Sanjay K., Ekta W (2011). Analysis of Electronic voting system in various countries. *International Journal on Computer Science and Engineering (IJCSE*. Vol. 3 No. 5.pp. 1925-1830.

Shamos, M. (2004). Paper v. Electronic Voting Records - An Assessment.

Sabo A., Siti J. & Rozita B. (2015) Issues and challenges of transition to e-voting Technology in Nigeria. Public Policy and Administration Research. Vol 5, pp 95-102.

Suleiman, A. T. & Gwani, Y. J. (2015). Mobile Electronic Voting System: Increasing Voter Participation. *JORIND*, Vol 13, ISSN 1596-8303. Available online at http://www.transcampus.org/JORINDV13DEC2015/Jorind%20Vol13%20No2%20Dec%20Chapter31.pdf [Accessed: 15 March 2015]

Vermont RD (2011). Australian ballot in the United States pp 38-39.

Idongesit Efaemiode Eteng nee Idongesit Fidelis Essien is a lecturer in Computer Science Department, University of Calabar, Nigeria, where she has taught since 2006. Prior to that, she worked for the same University as a programmer/ System's Analyst. She earned her PhD from the renowned University of Ibadan in 2013. Dr (Mrs) Eteng has supervised various research projects for both undergraduate and graduate students. She has also published several articles in both local and international journals. Her areas of interest are Internet Computing, Security, Distributed Databases, Optimization, Formal models in Software Engineering and health systems. She is a member of Computer professionals of Nigeria (CPN), Nigeria Computer Society (NCS) and IEEE. Dr (Mrs) Eteng lives in Calabar, Nigeria with her husband Pastor Efaemiode Eteng and children. She enjoys writing poems and short stories/books for children. She is also a public speaker and has co-authored a few practical manuals with her colleagues at the University of Calabar. Dr (Mrs) Eteng is the corresponding author and can be contacted at idongesitessien@yahoo.com, idongeteng@gmail.com and ideteng@unical.edu.ng.

Paul Umoren is a graduate of Computer Science Department, University of Calabar. He is a blogger and a Software Developer, and hangs out with the Camera during his spare time taking pictures of nature. He is a consultant to Girls' Power Initiative (GPI). His work experience includes software development at Axum Technologies; he is currently working for Partnership Opportunities for Women Empowerment Realization (POWER), a Non-governmental Organization, as a Communication/Information Technology Officer. He is single.

Ugochi Ahunanya held a Bsc. (Hons) degree in Computer Science from the University of Calabar, Calabar, Nigeria, until her death in the first quarter of this year, 2016. She was the best female student in her class. This paper is part of her project work which was supervised by Dr (Mrs) Eteng.

A Comparative Analysis of Feature Selection and Feature Extraction Models for Classifying Microarray Dataset

Arowolo M. Olaolu, Sulaiman O. Abdulsalam, Isiaka R. Mope, and Gbolagade A. Kazeem
Department of Computer Science, College of Information and Communication Technology, Kwara State University, Malete, Nigeria.
micheal.arowolo14@kwasu.edu.ng, abdulsalamny@gmail.com, imabdulrafiu@yahoo.com and kazeem.gbolagade@kwasu.edu.ng

ABSTRACT

Purpose: The purpose of this research is to apply dimensionality reduction methods to fetch out the smallest set of genes that contributes to the efficient performance of classification algorithms in microarray data.
Design/Methodology/Approach: Using colon cancer microarray dataset, One-Way- Analysis of Variance is used as a feature selection dimensionality reduction technique, due to its robustness and efficiency to select relevant information in a high-dimension of colon cancer microarray dataset. Principal Component Analysis (PCA) and Partial Least Square (PLS) are used as feature extraction techniques, by projecting the reduced high-dimensional data into efficient low-dimensional space. The classification capability of colon cancer datasets is carried out using a good classifier such as Support Vector Machine (SVM). The study is analyzed using MATLAB 2015.
Findings: The study obtained high accuracies and the performances of the dimension reduction techniques used are compared. The PLS-Based attained 95% accuracy having edge over the other dimension reduction methods (One-Way- ANOVA and PCA).
Practical Implications: The major implication of this research is getting the local dataset in the environments which lead to the usage of an open resource dataset.
Originality: This study gives an insight and implications of high dimensional data in microarray gene analysis. The application of dimensionality reduction helps in fetching out irrelevant information that halts the performance of a microarray data technology.

Keywords: *Dimension Reduction, One-Way-ANOVA, PCA, PLS, Classification.*
Paper Type: *Research work*

1 Introduction

Recently, procedures of machine learning in data mining have improved (Shengyan and Iagnemma, 2010; Techo, Nattee and Theeramunkong, 2008), microarray technology has become known as a proficient technique for cancer diagnosis, prognosis and treatment (Mukesh, Nitish, Amitav, and Satanu, 2015). Deoxyribonucleic Acid (DNA) microarray analysis have great effect by fetching informative genes that causes cancer (Flores, Hsiao, Chiu, Chuang, Huang and Chen, 2013), but its existing major drawback is caused by the curse of dimensionality, it hinders helpful information in a dataset which results into instability of computation. Therefore dimension reduction methods as a preprocessing step for retrieving similar data in a very high dimensional space is necessary, so as to select or extract relevant features which are very important in the analysis of microarray data of cancer (Mukesh, Nitish, Amitav, and Satanu, 2015).

To enhance the development of microarray technology (Mukesh, Nitish, Amitav, and Satanu, 2015), which has been supportive in studying thousands and millions of genes concurrently and also generating huge amount of data, several dimension reduction techniques and classifiers based on machine learning approach have been proposed in literature by various researchers and practitioners (Wang, Chu and Xie, 2007). Several algorithms and techniques have been proposed for dimensionality reduction. In feature selection, distinctive features are done by selection and

removal of redundant and irrelevant features (Cecille, Dana and Otman, 2013; Wald, Khoshgoftaar and Napolitano, 2013). Due to high dimensionality of original dataset, features are selected before classification, by using some feature selection approaches such Filter, wrapper and embedded methods as a dimension reduction method. Feature extraction, is one of the most important technique used in analysis and interpretation of microarray data, it is sectioned into supervised and unsupervised machine learning methods. Among the most effective methods are PCA, PLS and Independent Component Analysis (ICA) (Maryam and Mohammad, 2016), by considering the important components for classification.

Classification is an important aspect of cancer diagnosis and treatment; microarrays offer hope that cancer classification can be objective and highly accurate, providing clinicians with the information to choose the most appropriate forms of treatment. Classification operation performs intelligent discrimination by means of features obtained from dimensionality reduction methods. Many classifiers have been used for microarray analysis task, such as Fisher Linear Discrimination Analysis, Support Vector Machine, K-Nearest Neighbor (KNN) and aggregated classifiers (Atiyeh and Mohammad, 2016 ; Smitarani and Pratishya, 2014). In this paper, SVM is adopted based on its constructive learning procedures, it is used for classification tasks, and it uses linear models in implementing non-linear class boundaries by transforming input space using a non-linear mapping into a new space. SVM produces an accurate classifier with less over fitting and it is robust to noise (Vapnik, 1998).

In this paper, simple yet very efficient methods for cancer classification using dimension reduction methods are proposed. One-Way-ANOVA is used as a feature selection method to select a subset of the dataset and analyze the performance, PCA and PLS is used as feature extraction methods for dimensionality reduction, to project the dataset into a low-dimensional space and constructs a new dimension by analyzing relationships hidden in the dataset. Classification is carried out using SVM to classify the predetermined data. The experiment shows PLS outperforms PCA and One-Way-ANOVA by reducing the dimension and attaining a better performance.

2 Related Works

In literature, several studies have been carried out on the classification of the microarray data, using dimensionality reduction techniques and with the use of different classifiers. Vijayarani and Maria Sylviaa (2016) worked on two divisions of dimension reduction, Feature extraction (PCA, LDA) and feature selection (FA). Feature extraction techniques performed more adequate than the feature selection. Reduction was done to the larger medical dataset (Thyroid, Oesophagal) to decrease the curse of dimensionality. Ali, Paul and Mahdhu, (2011), Analyed microarraydata in the field of diagnosis and treatment of patients, a very high dimensional dataset containing some noise, non-useful information and a small number of relevant features for disease, proposes a non-linear dimensionality reduction algorithm Local Principal Component (LPC) which aims to maps high dimensional data to a lower dimensional space.

The reduced data represents the most important variables underlying the original data. Experimental results and comparisons were presented to show the quality of the proposed algorithm.

Research is still on for innovative techniques to select unique attributes or features so that the classification accuracy can be improved and the processing time can be reduced. From the above studies, there are limitations regarding to the problem of improving, developing efficiency and effectiveness of classification. Although a number of comparative studies have been made on feature selection, feature extraction and classification methods of gene expression of microarray data, they were all conducted on different gene expression data sets.

3 Methodology

3.1 Data Set

Colon cancer dataset was used in this experiment, it contains an expression of 2000 genes with highest minimal intensity across 62 tissues, derived from 40 tumor and 22 normal colon tissue samples (Alon, Barkai and Notterman, 1999). The gene expression was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. The gene intensity has been derived from about 20 feature pairs that correspond to the gene on the DNA microarray chip by using a filtering process. Details for data collection methods and procedures are described in (Alon, Barkai and Notterman, 1999), and the data set is available from the website http://microarray.princeton.edu/oncology/.

The computer configuration exploited for the purpose of comparative study uses iCore2 processor, 4GB RAM size, 64-bit System and MATLAB 2015a as the implementing tools.

3.2 Methods

This study presents an approach for classification of microarray data, which consists of two phases:

i. Input colon cancer dataset (Alon, Barkai and Notterman, 1999) and preprocess using dimension reduction methods such as One-Way-ANOVA as a feature selection technique, PCA and PLS as feature extraction techniques.
ii. After selecting the relevant features, SVM is applied to classify the microarray dataset in terms of accuracy, sensitivity, specificity and precisions (Fig.1).

To compute the importance of relevant gene in a dataset, One-Way-ANOVA based on p-value was used as a technique to analyze the colon cancer data, in which one or more response variables are measured under various conditions identified by one or more classification variables. In an analysis of variance, the variation in the response is separated into variation attributable to differentiate between the classifications variables and variation attributable to random error. An analysis of variance constructs tests to determine the significance of the classification effects. A typical goal in an analysis of variance is to compare means of the response variable for various combinations of the classification variables (Bharathi and Natatjan, 2010).

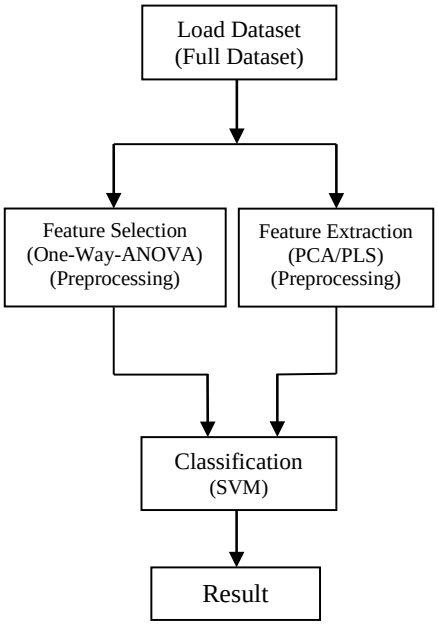

Figure 1: Technique Workflow

3.2 Dimensionality Reduction Methods

3.3.1 Feature Selection Method

The motivation is to carry out a one-way ANOVA feature with 0.05 p-values for selection of responsive data; it tests whether or not all the different classes of Y contain the equivalent mean as X, by adopting (Nadir, Othman and Ahmed, 2013). The following features apply:

$N_j = The\ number\ of\ cases\ with\ Y = j$
$X_j = The\ sample\ mean\ of\ predictor\ X\ for\ target\ class\ Y = j$

$S_j^2 = The\ sample\ variance\ of\ predictor\ X\ for\ target\ class\ Y = j$:

$$S_j^2 = \sum_{i=1}^{N_j}(X_{ij} - \bar{X}_j)^2/(N_j - 1) \quad (1)$$

$\bar{X}$: $The\ grand\ mean\ of\ predictor\ X$:

$$\bar{X} = \sum_{j=1}^{J} N_j X_j / N \quad (2)$$

The notations above are based on non-missing pairs of the sample and attribute of the dataset used in terms of (X, Y). The p-value is calculated by p value = Prob {F (J-1, N-J)>F}:
Where,

$$F = \frac{\sum_{j=1}^{J} N_j(x_j - \bar{x})^2/(J-1)}{\sum_{j=1}^{J}(N_j - 1)x_j^2/(N-1)} \quad (3)$$

F (J-1, N-1) is an indiscriminate variable that follows an F distribution with level of freedom J-1 and N-J. If the denominator for a predictor is zero, set the p-value = 0.5 for the predictor. Predictor is classed by sorting according to the p value in ascending order. If there is tie, sort by F in descending order and if it still ties sort by N in descending order. Classification of features shows that out of 2001 features, 416 features were the most significant features correlated to the microarray data analysis. It uses the p-values to rank the significant features with small values and the sorted numbers of features are further processed.

3.3.2 Principal Component Analysis (PCA)

PCA is a widely used unsupervised feature extraction technique; it works by replacing the original variables in a data with numerical variables called principal component by capturing the most descriptive features with respect to the most relevant ones (Arif, 2009). PCA mathematically transforms data by referring them to a different coordinate system in order to obtain the greatest variance. A number of correlated variables into a smaller number of uncorrelated variables called principal components (Smitarani and Pratishya, 2014).

PCA identifies patterns of similarities and differences in a data, these patterns are determined and can be compressed by reducing the numbers of dimensions without much loss of information.
In order to conduct the PCA analysis for the input data, the following steps are performed by adopting [16]:

1) Create N x d data matrix with one row vector $\boldsymbol{x_n}$ per data input
2) Subtract mean from each row vector $\boldsymbol{x_n}$ in X
3) Calculate the covariance matrix of X
4) Find Eigen Vectors and Eigen values of $\sum$
5) Fetch the Eigen vector with the largest Eigen values

The PCA are uncorrelated and the components explain the largest percentage in the dimensional dataset with results in extracting 10 components which are considered relevant in the colon cancer dataset used.

3.3.3 Partial Least Square (PLS)

Partial Least Square (PLS) is a supervised feature extraction technique, which is widely used as a procedure in modeling associations linking blocks of experimental variables by means of latent variable, it tries finding uncorrelated linear transformations (latent components) of the original predictor variables which have high covariance with the response variables (Xue, and Guo, 2014). The goal of PLS is to find the linear relationship between the response and explanatory variables y and X:

$$X = TP^T + E_x \quad (4)$$

$$y = TC^T + E_y \quad (5)$$

Where T represents the scores (latent variables) P and C are loadings, and $\boldsymbol{E}x$ and $\boldsymbol{E}y$ are the residual matrices obtained the original **X** and **y** variables.

Feature extraction using PCA ignores the response variable and its equivalence. PLS integrates the response variable during the dimensionality reduction procedure. PLS outperforms PCA in the case of microarray gene expression, PLS only consists of indicating the amount of gene components whereas PCA necessitates choosing the essential gene

components (Nebu, Vinjay, Gayathri and Jaisankra, 2012).

3.4 Classification

In this step, the results for classification are computed using SVM for classification. SVM is a constructive learning procedure based on statistical knowledge theory (Vapnik, 1998), it is used for classification tasks, and it uses linear models in implementing non-linear class boundaries by transforming input space using a non-linear mapping into a new space. SVM produces an accurate classifier with less over fitting and it is robust to noise.

Assuming $\{(x_1, y_1), \dots, (x_n, y_n)\}$ be a training set with $x_{1i} \epsilon R^d$ and y_i is the corresponding target class. SVM can be reformulated as:

Maximize:

$$\mathbf{J} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T, x_j) \quad \mathbf{(6)}$$

Subject to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \; and \; \alpha_i \geq 0, i = 1, 2, \dots, n \quad \mathbf{(7)}$$

This is the weighted average of the training features. Here, αi is a Lagrange multiplier of the optimization task and αi is a rank label. Values of α′si are non-zero for all the points lying inside the margin and on the correct side of the classifier. The kernel function is used to solve the problem. The Kernel function analyses the relationship among the data and it creates a complex divisions in the space (ML, 2015).

4 Results, Data Analysis and Discussion

To assess the performance of the proposed approach, several experiments have been conducted on a publicly available dataset (Alon, Barkai, and Notterman, 1999). A brief description of the dataset (the salient features of each dataset is summarized in Table I).

Table I: Result Evaluations

Dataset	Features Selected (one-way ANOVA)	Feature Extracted (PCA)	Feature Extracted (PLS)
Colon Cancer (2001x62)	416	10 Components	20 Components

In the first experiment, feature selection methods using One-Way-ANOVA and SVM as the classifier for the function of the number of genes retained: 416 out of 2001. The average performance of the approach across the dataset (the performance for the dataset is reported in a confusion matrix).

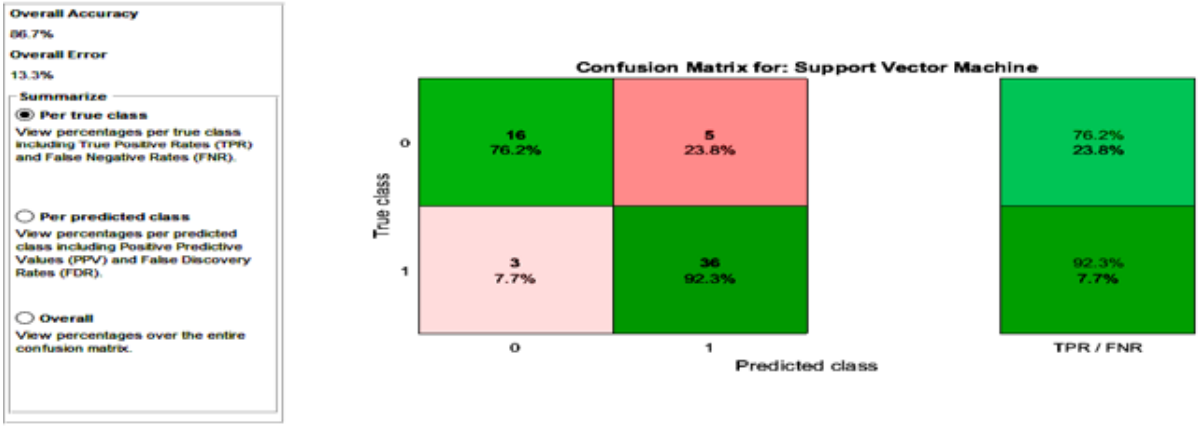

Figure 2: Confusion Matrix of Proposed Classification, using One-Way-ANOVA for Classification

True Positive Rate 76.2% and False Negative Rate yields 92.3%. TP=36 FP=5 FN=3 TN=16
ACCURACY: *(TP + TN) / (TP+TN+FP+FN) = 86.67%*
SENSITIVITY: *TP/ (TP+FN) = 92.31*
SPECIFICITY: *TN/ (FP+TN) = 76.19*
PRECISION: *TP/ (TP+FP) = 87.81*

In the second experiment, feature extraction method is compared using PCA to reduce the high-dimension and SVM is used as the classifier. PCA is used to de-correlate the data and 10 components was achieved, in Fig. 3, the overall accuracy on all the datasets obtained using PCA as feature extraction to transform and extract the dataset is reported in a confusion matrix.

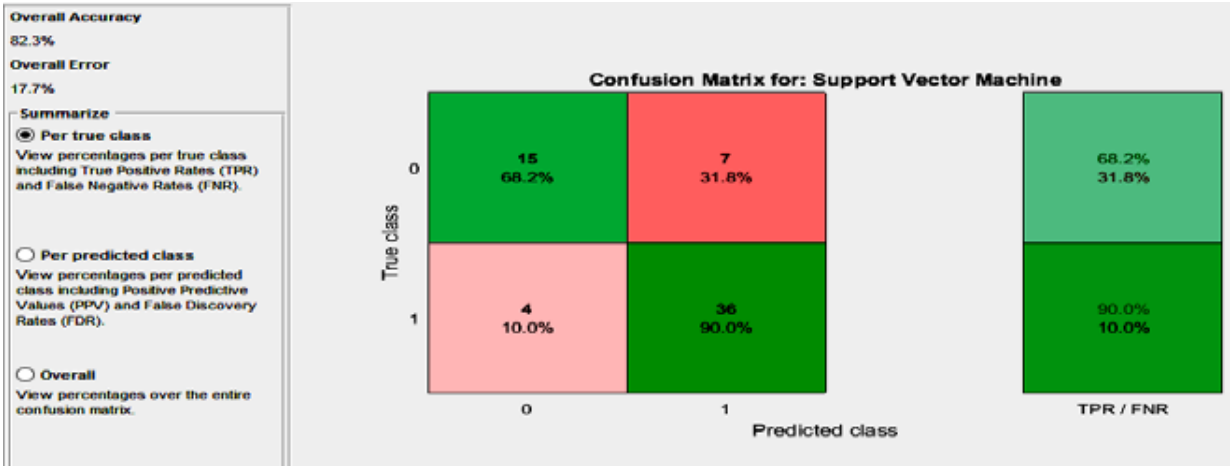

Figure 3: Confusion Matrix of Proposed Classification, using PCA-Based for Classification

True Positive Rate 68.2% and False Negative Rate yields 90.0%. TP=36 FP=7 FN=4 TN=15
ACCURACY: *(TP + TN) / (TP+TN+FP+FN) = 82.3%*
SENSITIVITY: *TP/ (TP+FN) = 90.00*
SPECIFICITY: *TN/ (FP+TN) = 68.18*
PRECISION: *TP/(TP+FP) = 83.72*

In the third experiment, this paper evaluates the performance of colon cancer dataset using PLS as a feature extraction method with SVM as a classifier. It obtained a better overall accuracy compared to the first and second experiment as shown in the confusion matrix of Fig. 4, it de-correlated the features of colon cancer dataset into 20-components which are considered relevant.

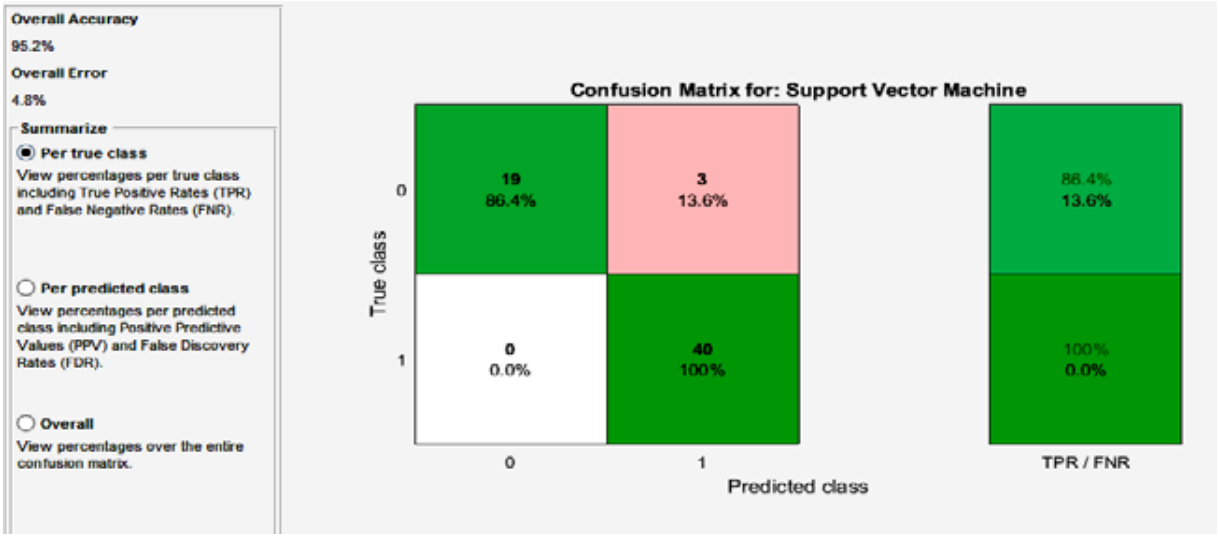

Figure 4: Confusion Matrix of Proposed Classification, using PLS for Classification

True Positive Rate 86.4% and False Negative Rate yields 100.0%. TP=40 FP=3 FN=0 TN=19
ACCURACY: *(TP + TN) / (TP+TN+FP+FN) = 95.16%*
SENSITIVITY: *TP/ (TP+FN) = 100*
SPECIFICITY*: TN/ (FP+TN) = 86.36*
PRECISION: *TP/ (TP+FP) = 93.02*

Table II: Performance Evaluation of Proposed One-Way-ANOVA, PCA and PLS Methods

S/No	Performance Metrics	One-Way-ANOVA Method	PCA Method	PLS Method
1	Training Time	71.2777	47.2686	5.0088
2	Accuracy (%)	86.67	82.30	95.16
3	Sensitivity (%)	92.31	90.00	100
4	Specificity (%)	76.19	68.18	86.36
5	Precision (%)	87.81	83.72	93.02
6	Area Under Curve	0.907204	0.848864	0.994318
7	Error (%)	13.3	17.7	4.8

Table II illustrates a comparative chart between the three methods used in terms of several performance measures such as accuracy, sensitivity, specificity, precision, error, time and area under curve. This comparison shows the integrity of the proposed approach with respect to the state of the art. The colon cancer dataset used to generate our result achieved its best on PLS for feature extraction; it makes this method suitable for practitioners.

5 Conclusion

In this paper, a widely used datasets was used in the literature for the evaluation of the algorithms used. The dimension reduction algorithms used to eliminate high dimensional data were One-Way-

ANOVA, PCA and PLS, it uses SVM as its classifier, and it was successfully implemented on MATLAB. For the purpose of finding the smallest gene subsets for accurate cancer classification, PLS method is highly effective compared to One-Way-ANOVA and PCA.

References

Shengyan, Z., & Iagnemma, K. (2010). Self-supervised learning method for unstructured road detection using Fuzzy Support Vector Machines, *International Conference on Intelligent Robots and Systems*, IEEE, pp. 1183–1189.

Techo, J., Nattee, C., & Theeramunkong, T. (2008). A Corpus-Based Approach For Keyword Identification Using Supervised Learning Techniques, *5th International Conference On Electrical Engineering/Electronics, Computer, Telecommunications And Information Technology*, Ecti-Con, Vol 1, pp. 33–36.

Mukesh, K., nitish, K.R., Amitav, S., & Santanu, K.R. (2015). Feature Selection and Classification of Microarray Data Using Map Reduce Based ANOVA and K-Nearest Neighbor, *11th International Multi-Conference on Information Processing, Procedia Computer Science*, Vol 54, pp. 301-310.

Flores, M., Hsiao, T., Chiu, Y., Chuang, E., Huang, Y., & Chen, Y. (2013). Gene Regulation, Modulation, and their Applications in Gene Expression Data Analysis, *Advances in Bioinformatics* 2013, pp. 360678–360678.

Wang, L., Chu, F., & Xie, W. (2007). Accurate Cancer Classification using Expressions of very Few Genes, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (TCBB), Vol 4, No 1, pp. 40–53.

Cecille, F., Dana, K., & Otman, B. (2013). Feature Selected Tree-Based Classification, *IEEE Transactions on Cybernetics*, Vol 43, No. 6, pp. 199-204.

Wald, R., Khoshgoftaar, T.M., & Napolitano A. (2013). Stability of Filter- and Wrapper-Based Feature Subset Selection, *IEEE 25th International Conference On Tools With Artificial Intelligence*, pp. 374–380.

Maryam, M., and Mohammad, H.M. (2016). *A Novel Feature Extraction Approach Based on Ensemble feature Selection and Modified Discriminant Independent Component Analysis for Microarray Data Classification*, Nałęcz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. Published by Elsevier, Vol 36, pp. 521-529.

Atiyeh, M., & Mohammad, H.M., (2016), Robust Feature Selection from Microarray data Based on Cooperative Game Theory and Qualitative Mutual Information. *Advances in Bioinformatics*, Hindawi, Article ID 1058305, 16 pages.

Smitarani, S., & Pratikshya, M., 2014 “Microarray Classification Using Intelligent Techniques”, International Journal of Scientific and Engineering Research, Vol 5, No. 7, pp. 1-5

Vapnik, V.N. (1998). Statistical Learning Theory. New York: John Wiley & Sons.

Alon, U., Barkai, N., & Notterman, D.A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceeding of Nationall myAcademic Sciences USA 1999, Vol 96, pp. 6745-50

Bharathi, A., & Natatjan, A.M. (2010), Cancer Classification of Bioinformatics using ANOVA, International Journal of Computer Theory and Engineering, Vol 2, No. 3, pp. 369-373.

Nadir, O.E., Othman, I., & Ahmed, H.O. (2013). A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. Research Journal of Applied Sciences, Engineering and Technology Vol 7, No. 3, pp. 625-638.

Arif, M.W. (2009), Microarray Classification with Hybrid Approaches, *Canadian Journal of Pure and Applied Sciences*, Vol 3, No. 1, pp. 759-763.

Vijayarani, S., & Sylviaa, S.M., (2016), Comparative Analysis of Dimensionality Reduction Techniques, *International Journal*

of Innovative Research in Computer and Communication Engineering, Vol 4, No 1, pp. 23-29.

Xue-Qjang, Z., & Guo-Zheng, L., (2014), Dimension Reduction for Protein Recognition by Using Incremental Partial Least Squares, *IEEE*. pp. 73-79.

Nebu, V., Vinay, V., Gayathri, P., & Jaisankra, N. (2012). A Survey of Dimensionaliity Reduction and Classification Methods. *IJCSES*, Vol 3, No. 3, pp. 45-54.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: John Wiley & Sons.

Vijayarani, S. & Maria, S. S (2016). Comparative Analysis of Dimensionality Reduction Techniques, *International Journal of Innnovative Research in Computer and Communication Engineering*, Vol 4, No 1, pp. 23-29.

Ali, A., Paul J.K., & Madhu, G. (2011). Dimension Reduction of Microarray Data Based on Local Principal Component, *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* Vol 5, No 5.

About the authors

Micheal O. Arowolo Holds his B.Sc and M.Sc degree in Computer Science from Al-Hikmah University, Ilorin, and Kwara State University, Malete, Nigeria respectively. He is an Oracle Certified Expert, a member of the IAENG, EAI and SDIWC. His research interests are Bio-informatics, Datamining, Machine Learning and Software Engineering.

Email: micheal.arowolo14@kwasu.edu.ng

Sulaiman O. Abdulsalam He holds a Bachelor of Science degree and Master of Science Degree in Computer Science, both from University of Ilorin, Nigeria. He is currently a lecturer at the Department of Computer Science, Kwara State University, Malete, Nigeria. He is a member of Nigeria Computer Society. His research intrests include Data Mining, Machine Learning and Software Engineering.

Email: abdulsalamny@gmail.com

Rafiu M. Isiaka he has his Ph.D. in Computer science, he is a lecturer in the Department of Computer Science, Kwara State University Malete since 2009. His research interest includes soft computing, e-learning, data mining and information security.

Email: imabdulrafiu@yahoo.com

Kazeem A. Gbolagade is a Professor and Provost at the College of Computer in Information Science, Kwara State University, Malete, Nigeria. He was born in Iwo (Osun State), Nigeria, on the 27th of August, 1974. He received his B.Sc degree in 2000 in Computer Science from the University of Ilorin, Kwara State, Nigeria. In 2004, he obtained his Masters degree from the University of Ibadan, Nigeria. In April 2007, he joined the Computer Engineering Laboratory group at the Delft University of Technology (TU Delft), The Netherlands. In TU Delft, he pursued a PhD degree under the supervision of Prof. Sorin Cotofana. He is a member of the IEEE. His research interests include Digital Logic Design, Computer Arithmetic, Residue Number Systems, VLSI Design, and Numerical Computing.

Email: kazeem.gbolagade@kwasu.edu.ng