Breast Cancer Surgery Survivability Prediction Using Bayesian Network and Support Vector Machines

Dania Abed Aljawad[1], Ebtesam Alqahtani[2], Ghaidaa AL-Kuhaili[3], Nada Qamhan[4], Noof Alghamdi[5], Saleh Alrashed[6], Jamal Alhiyafi[7], Sunday O. Olatunji[8] (corresponding author)

Computer Science Department, College of Computer Science and Information Technology, University of Dammam, Dammam, Kingdom of Saudi Arabia

[1]danya.abdaljawad@gmail.com, [2]ebtesam.j.alqahtani@gmail.com, [3]ghaidaa_alkuhaili@yahoo.com, [4]nada.qamhan@gmail.com, [5]noof.abdullah.alghamdi1993@gmail.com, [6]saalrashed@uod.edu.sa, [7]jalhiyafi@uod.edu.sa, [8]oluolatunji.aadam@gmail.com, [8]osunday@uod.edu.sa

***Abstract*— Predicting the survival status of patients who will undergo breast cancer surgery is highly important, where it indicates whether conducting a surgery is the best solution for the presented medical case or not. Since this is a case of life or death, the need to explore better prediction techniques to ensure accurate survival status prediction cannot be overemphasized. In this paper we evaluate the performance of support vector machine (SVM) and Bayesian network (BN) in predicting the survival state of breast cancer patients after having a surgery. The experiments on both techniques have been carried out using Weka software package. Empirical results from simulations showed that support vector machine outperformed Bayesian network in this task, where support vector machine achieved better accuracy of 74.44% while Bayesian network had its best accuracy of 67.56%.**

Keywords—Haberman's survival dataset; Bayesian network; support vector machine; machine learning; supervised learning.

I. INTRODUCTION

Machine learning is one of the recent and most important fields of artificial intelligence [1]. It is a very powerful tool to solve complex real life problems in medical, business, and educational fields [1]. Studying and evaluating the performance of machine learning techniques in the medical field helps draw important findings contributing to the enhancement of others' lives. Considering surgery to cure breast cancer is one of the most critical medical decisions, because such a decision will significantly affect the patient's survival status. Since this decision involves a case of life or death, the need to find better prediction techniques to ensure accurate survival status prediction is imperative. In contributing towards this important medical decision making process, we have selected the popular Haberman's Survival Dataset to evaluate the performance of Support Vector Machine and Bayesian Network classification algorithms in predicting the survival status of patients who will undergo breast cancer surgery. SVM and BN have been selected due to their reputation of achieving accurate results in several fields of applications. The experimental settings and simulations have been carried out using the popular machine learning and data mining software package known as Weka [2].

SVM is very well known for its good performance even in the case of having small datasets [3]. Hence, it is considered to be valuable to explore its unique ability on a relatively small dataset like Haberman's Survival dataset. Additionally, BN is a very powerful tool to represent causal relationships between different variables using conditional probability [4]. Thus, we want to study its effectiveness in classifying the instances of the selected dataset which will give an indication whether the features are conditionally dependent or not. Such an evaluation can reveal the strengths and weaknesses of each technique in the selected field.

There are various studies that were conducted on Haberman's survival dataset using different machine learning techniques. Such works include, "Learning and classification with prime implicants applied to medical data diagnosis" [5], "A two-step supervised learning artificial neural network for imbalanced dataset problems" [6], "Comparative study of advanced classification methods" [7], "Classification of imbalanced dataset using conventional Naïve Bayes classifier" [8], and "An Experimental Analysis of Clustering Algorithms in Data Mining using Weka Tool" [9]. However, none of these papers compares the performance of BN and SVM techniques. Hence, conducting such a study is considered to be important revealing the potential strengths and weaknesses of these two popular techniques in the selected field, and contributing to facilitating pre-surgery decision making process.

Empirical results emerging from the conducted experiment showed that SVM outperformed BN in this task, where SVM achieved its best accuracy of 74.44% while BN achieved its best accuracy of 67.56%.

The remaining part of this paper is organized as follows. Section II provides a description of the evaluated machine learning techniques in this paper. Section III contains

978-1-4673-8765-1/17/$31.00 ©2017 IEEE

empirical studies that include dataset description, experimental setup, and the adopted optimization strategy. Section IV presents the results and discussion emanating from the studies conducted. Finally, section V contains the conclusion and recommendation emerging from this work.

II. Description of the Evaluated Techniques

A. Support Vector Machine Technique

SVM is a supervised machine learning algorithm which can be used for analyzing data, classification, and regression analysis [10].

SVM can be divided into Linear SVM and Non-Linear SVM. Linear SVM is used to classify data points in a feature space into two classes if they are linearly separable, i.e., they can be separated using a line, Fig. 1 illustrates this type of data points [11].

Linearly separable datasets are actually well handled using linear classifiers. However, linear classifiers are not advanced enough to handle other complex non-separable datasets. As can be seen from Fig. 2, data points in non- separable datasets cannot be separated using a line which means that using linear classifiers is not an appropriate method to solve this problem [12]. Therefore, Isabelle Guyon, Bernhard Boser, and Vladimir Vapnik proposed nonlinear classifier as a solution for this problem. Nonlinear SVM classifier maps the dataset to a higher dimensional space that includes nonlinear features. Then, it finds the maximal margin hyperplane that linearly separates the data in the new feature space [13], as can be seen in Fig. 3. However, if you look at the left side of Fig. 4, you can see that the mapping process will be hard in such a case. This process can use several or infinite dimensions which makes it hard or sometimes impossible to compute. Most of the times, the mapping process requires intensive computations, which makes it less efficient, and this is absolutely not desirable [14]. Fortunately, kernel function helps to solve this problem. Using kernel functions, one can avoid calculating explicit features values in the higher dimensional space and rely on simpler inner dot product operations instead [12].

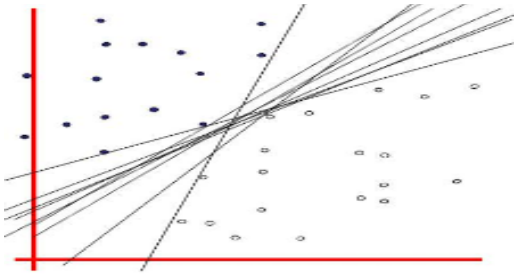

Fig. 1. Linearly separable data [11].

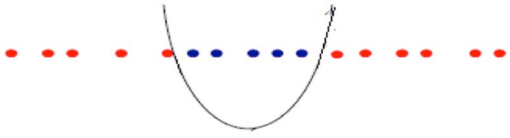

Fig. 2. Nonlinearly separable dataset [12].

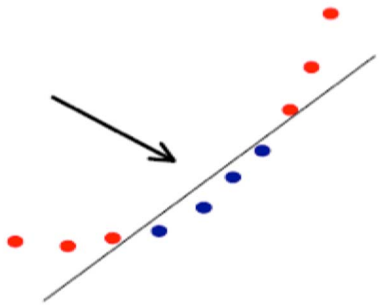

Fig. 3. Separating nonlinearly separable data in a higher dimensional space [12].

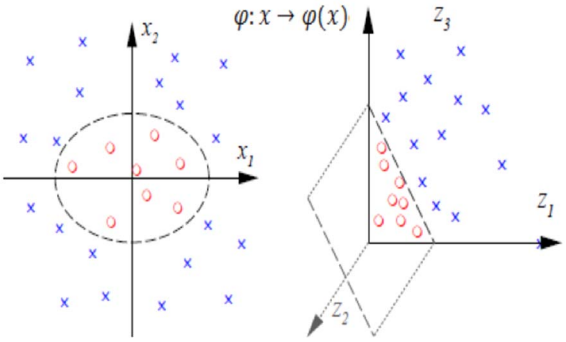

Fig. 4. Mapping process [15].

B. Bayesian Network Technique

Bayesian network which is also known as belief network is one of the most powerful machine learning techniques. It is an integration of various principles from computer science, graph theory, statistics, and probability theory [16].

BN represents knowledge as graphical forms or structures, because of that BN is considered as a member of the probabilistic graphical models family (models that express conditional dependencies between random variables) [16].

BN graph consists of a set of nodes that represent random variables connected by directed edges or links that represent the probabilistic or conditional dependencies between the connected nodes (variables). These probabilistic dependencies are estimated by the use of known statistical and computational techniques. BN graph is a directed acyclic graph (DAG), i.e., its links are directed and there are no cycles between its nodes.

A directed acyclic graph consists of nodes that are represented as circles labeled by the variables names and connected by directed links or edges that represent the dependencies between the variables. Fig. 5 shows a simple example of DAG.

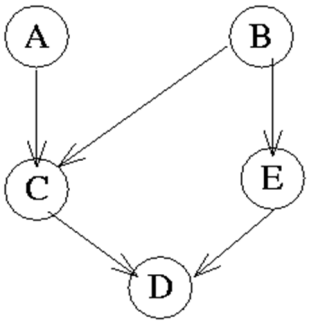

Fig. 5. Directed acyclic graph.

Each variable in BN is independent of all other non-descendant variables given its parents' states. This property reduces the parameters needed to describe the joint probability distribution (JPD) of every variable [16].

BN uses DAG to represent the qualitative part of the model while the quantitative part of the model is represented by a way that is consistent with Markovian property, where the conditional probability distribution (CPD) of any node is calculated based on its parent nodes only [16].

III. Empirical Studies

A. Description of the Dataset

Haberman's survival dataset contains information about the survival of the University of Chicago's Billings Hospital's patients who had undergone surgery for breast cancer. This dataset consists of 306 samples with four numerical attributes which are patient's age, patient's year of operation, number of positive nodes detected, and patient's survival status that illustrates whether he/she had survived after the operation or not. This dataset does not contain any missing values. 225 instances of the 306 instances belong to class '1' which represents the patients who had survived for 5 years or longer after the surgery. The other 81 instances belong to class '2' which represents the patients who died within 5 years after the surgery.

1) Statistical Analysis of the Dataset

The statistical analysis of the dataset is presented in table I. The mean, median, standard deviation, maximum, and minimum values of each feature in the dataset are presented. Also, we conducted a correlation analysis between every feature and the class variable to have a clear insight into how they are related. The correlation coefficient obtained from the conducted correlation analysis is shown in table II.

TABLE I. Statistical Analysis of the Dataset

Statistical Measure	Age Of Patient	Year Of Operation	Number Of Positive Nodes Detected
Mean	52.458	62.853	4.026
Median	52	63	1
Standard deviation	10.803	3.249	7.19
Maximum	83	69	52
Minimum	30	58	0

TABLE II. The Correlation between Each Attribute and the Target Attribute

Features' Pairs	Correlation Coefficient
Age and survival status	0.06795
Year and survival status	-0.00477
Positive nodes and survival status	0.28677

B. Experimental Setup

We carried out our experiment on the selected dataset using Weka. Since we are examining a classification problem, the class label is expected to be nominal not numerical. Although the class variable in the selected dataset (survival status) contains numerical values, it should be considered as nominal since we do not apply any mathematical operations on such values. Rather, we only use them as class labels. Since the class variable consists entirely of numbers, Weka considered it as a numerical target variable which conflicts with the purpose of the classification process. Hence, we first converted the values of the class variable into nominal values by applying the unsupervised attribute level filter known as 'NumericToNominal' on the class variable.

Before performing the classification process, it should be noted that Haberman's survival dataset is imbalanced where there are 306 instances, 225 of them have the class label '1' (the patient survived for 5 years or longer), whereas the remaining 81 instances have the class label '2' (the patient died within 5 years). This imbalance can affect the accuracy of the classification model to be generated. In other words, the classification model can get biased to the majority of the instances with a specific class label resulting in a misclassification of the instances which belong to the minority class label. To solve this problem, we used Synthetic Minority Oversampling Technique (SMOTE) filter which synthetically generates more instances that belong to the minority class label in order to balance the dataset. Since there are 225 instances with the class label '1', and 81 instances with the class label '2', setting the percentage of generating the synthetic instances to be 178% of the total number of the minority class instances is a proper choice. This is because such a percentage will result in generating 144 more instances that belong to class '2' leading to having a total of 225 instances in this class, which results in having the same number of instances for every class label (225). We kept the

seed of SMOTE as 1, and the number of nearest neighbors needed to synthesize an instance from another one as 5.

C. Optimization Strategy

1) Support Vector Machine Optimization Process

There is a set of parameters that must be used to conduct the classification process using SVM, which are cost (C), hyper-parameter (Lambda), epsilon (ε), and kernel function. These parameters have to be optimized to get an optimal classification result. We used a greedy approach to optimize these parameters, where we fixed other parameters' values while trying to optimize the current one to get the highest classifier accuracy at the current step. For example, while trying to optimize the selection of C, we kept the values assigned to other parameters as they are, and we only kept searching for an optimal value for C. If the other parameters were not yet optimized while optimizing a specific parameter, we kept the default values assigned to them by Weka, otherwise we kept their optimal values.

We optimized the parameters in the following order, kernel function, cost, Lambda, and epsilon. Defining the kernel function first is important because it establishes the bases on which other parameters will be defined. During the optimization process we used 10-fold cross validation to evaluate the classifier accuracy. The optimal SVM parameters obtained as a result of applying the optimization process are listed in table III. These parameters gave the highest accuracy of SVM when applied on the selected dataset, and the obtained accuracy was 74.44%.

2) Bayesian Network Optimization Process

BN classifier is another classifier that we used to classify the selected dataset after optimizing its parameters. There are two main parameters that control how BN classifier works, which are, search algorithm parameter and estimator parameter. As in the case of optimizing the parameters of SVM, we used the greedy approach in optimizing BN parameters. We optimized the search algorithm parameter first and then we used this optimized search algorithm value to optimize the second parameter. The parameters of the optimal BN model are illustrated in table III. The optimal BN model had an accuracy equal to 67.56% when applied on the selected dataset.

TABLE III. THE OPTIMAL PARAMETERS OF SVM MODEL AND BN MODEL

Model	Parameters	Optimal Value Chosen
SVM	C	27
	Hyper-parameter (Lambda)	0.3
	Epsilon (ε)	0.001
	Kernel	Gaussian
BN	Search Algorithm	K2
	Estimator	SimpleEstimator

IV. RESULT AND DISCUSSION

We used confusion matrices, Receiver Operator Characteristic (ROC) curves, recall, and precision of SVM and BN optimal models to further investigate their performance. As can be seen from the confusion matrix of the SVM classifier in Fig. 6, 59 of the patients who survived 5 years or longer (class 1) were mistakenly classified as patients who died within 5 years (class 2). On the other hand, 56 of the patients who died within 5 years (class 2) were mistakenly classified as patients who survived for more than 5 years (class 1). The ROC curve of any classifier describes its testing accuracy with respect to positive instances. It measures the change rate between false positives (x axis) and true positives (y axis) of a specific class label, the greater the area under the curve the better the classification is. Excellent classifiers should have an area under ROC curve ranging from 0.9 to 1, while good classifiers should have an area under ROC curve ranging from 0.8 to 0.9 [17]. Fig. 7 and Fig 8 depict the ROC curve of the final SVM model with respect to class 1 and class 2 respectively. As can be seen from Fig. 7 and Fig. 8, the area under the ROC curve for both classes is 0.7444, which indicates that the classifier accuracy is not bad yet not so good.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 166  59 |   a = 1
  56 169 |   b = 2
```

Fig. 6. Confusion matrix of SVM final model.

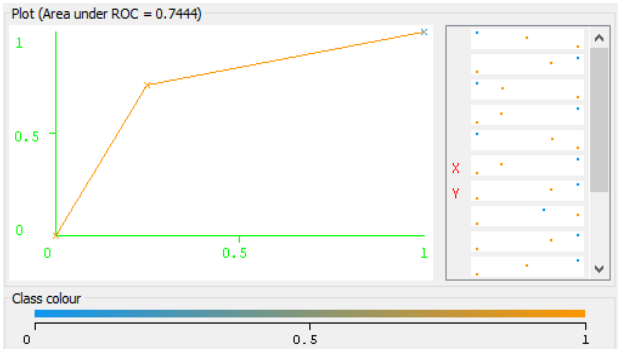

Fig. 7. ROC curve for class 1 of SVM final model.

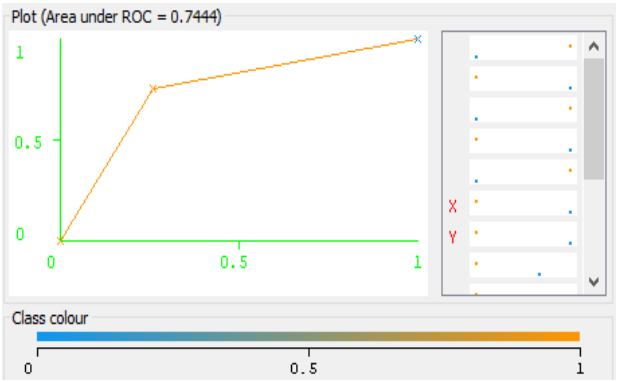

Fig. 8. ROC curve for class 2 of SVM final model.

Fig. 9 shows the confusion matrix of the final BN model. As can be seen from the confusion matrix in Fig. 9, 49 of the patients who survived 5 years or longer (class 1) were mistakenly classified as patients who died within 5 years (class 2). On the other hand, 97 of the patients who died within 5 years (class 2) were mistakenly classified as patients who survived more than 5 years (class 1). We have visualized this classification result using ROC curves for both classes. Fig. 10 and Fig. 11 show the ROC curves of class 1 and class 2 respectively. As the figures show, the area under the ROC curves for both classes is 0.6753, which indicates that the classifier accuracy is not good.

Additionally, both classifiers were further evaluated based on recall and precision. BN had a higher value of recall indicating more reliability in finding positive class instances (class 1) among actual positive class instances. On the other hand, SVM had a greater value of precision emphasizing the fact that SVM was more capable of finding positive class instances among relevant and irrelevant positive class instances. Table IV compares both classifiers based on accuracy, recall, and precision.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 176  49 |   a = 1
  97 128 |   b = 2
```

Fig. 9. Confusion matrix of BN final model.

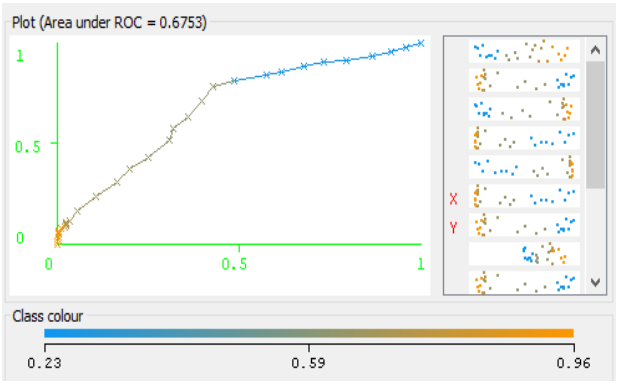

Fig. 10. ROC curve for class 1 of BN final model.

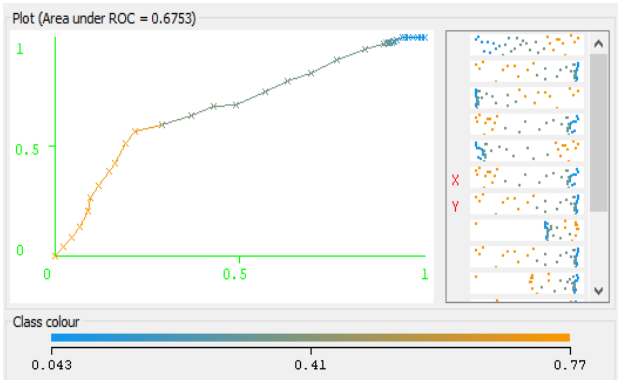

Fig. 11. ROC curve for class 2 of BN final model.

TABLE IV. A COMPARISON BETWEEN THE CLASSIFIERS' PERFORMANCE

Classifier	Accuracy	Recall	Precision
SVM	74.44%	73.78%	74.77%
BN	67.56%	78.22%	64.47%

A. *Further Discussions*

As can be seen from the conducted experiment, SVM outperformed BN in classifying the instances of the selected dataset in spite of optimizing both classifiers using the greedy approach discussed earlier. This indicates that the dataset is not suitable to be classified using BN, as it seems that its features are not highly dependent on each other. Additionally, this result proves the well-known property of SVM which is outperforming other algorithms when applied on small datasets, where it performed better than BN in this experiment.

V. CONCLUSION AND RECOMMENDATION

Two models for predicting the survivability state of patients who will undergo breast cancer surgery have been proposed in this work using SVM and BN classifiers. Empirical results from the conducted experiments clearly indicated that SVM performed better than BN when applied on Haberman's survival dataset. The results encourage conducting more studies to evaluate the performance of SVM in the selected field and come up with different ways to increase its resulting accuracy.

It should be stated that Haberman's survival dataset has a limited set of features, which may not be enough to evaluate the survival status of breast cancer patients. Thus, future work should consider testing the used techniques on a dataset with more features like the stage of the cancer, the use of radiotherapy, and the use of chemotherapy treatment. This can help improving the classification performance and matching the latest cancer research outcomes in terms of new features discovered recently.

Additionally, we recommend evaluating the performance of the used techniques before and after using SMOTE filter to investigate the effect of balancing the dataset on the classification performance.

Also, further studies could be conducted in the direction of investigating the effect of feature selection and other possibilities on the performance of the techniques proposed. The promising nature of these techniques could be an impetus to further explore their prediction capabilities in other health informatics related tasks.

Acknowledgment

The University of Dammam is hereby acknowledged for some of the facilities utilized during the course of this research.

References

[1] Das, S., Roy, N., Dey, A., & Pal, A. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect, 115 (9), 31-41. Retrieved from http://research.ijcaonline.org/volume115/number9/pxc3902402.pdf

[2] Downloading and installing Weka. Retrieved April 1, 2016, from http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[3] Goebel, R. (2014). Support Vector Machines (SVMs). Retrieved April 4, 2016, from http://www.brainvoyager.com/bvqx/doc/UsersGuide/MVPA/SupportVectorMachinesSVMs.html

[4] Friedman, N. & Koller, D. (2003). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. Journal of Machine Learning, 50 (1), 95-125. http://doi.org/10.1023/A:1020249912095

[5] Shevked, Z., & Dakovski, L. (2007). Learning and classification with prime implicants applied to medical data diagnosis. Proceedings of the 2007 international conference on Computer systems and technologies - CompSysTech '07. http://doi.org/10.1145/1330598.1330708

[6] Adam, A., Ibrahim, Z., Shapiai, M. I., Chew, L. C., Jau, L. W., Khalid, M., & Watada, J. (2012). A two-step supervised learning artificial neural network for imbalanced dataset problems. International Journal of Innovative Computing, Information and Control, 8(5 A), 3163–3172. Retrieved from http://www.ijicic.org/contents.htm

[7] Shruti, A., & B. I., K. (2015). Comparative Study of Advanced Classification Methods. International Journal on Recent and Innovation Trends in Computing and Communication, 3(3), 1216-1220. http://doi.org/10.17762/ijritcc2321-8169.150371

[8] Sobran, N., Ahmad, A., & Ibrahim, Z. (2013). Classification of Imbalanced Dataset Using Conventional Naïve Bayes Classifier. International Conference on Artificial Intelligence in Computer Science and ICT, 35-42. Retrieved from http://worldconferences.net/proceedings/aics2013/toc/papers_aics2013/A021 - NUR MAISARAH MOHD SOBRAN - Classification of Imabalanced dataset using conventional naive bayes classifier.pdf

[9] Goyal, V. K. (2014). An Experimental Analysis of Clustering Algorithms in Data Mining using Weka Tool. International Journal of Innovative Research in Science & Engineering, 2 (4), 171-176. Retrieved from http://ijirse.in/docs/Apr14/IJIRSE140406.pdf

[10] Introduction to Support Vector Machines. Retrieved February 12, 2016, from docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[11] Jakkula, V. (2006). Tutorial on Support Vector Machine (SVM). School of EECS, Washington State University. Retrieved from http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf

[12] Berwick, R. (2003). An Idiot's guide to Support vector machines (SVMs). Retrieved from https://web.cs.dal.ca/~tt/CSCI415009/Berwick03.pdf.

[13] Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. The Analyst, 135(2), 230–267. http://doi.org/10.1039/b918972f

[14] Cristianini, N. (2001). Support Vector and Kernel Machines. BIOwulf Technologies. Retrieved from http://www.support-vector.net/icml-tutorial.pdf

[15] Weston, J. Support Vector Machine (and Statistical Learning Theory) Tutorial. NEC Labs America 4, Independence Way, Princeton, USA. Retrieved February 11, 2016 from http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf

[16] Ben-Gal, I. (2008). Bayesian Networks. Encyclopedia of Statistics in Quality & Reliability. http://doi.org/10.1002/9780470061572.eqr089

[17] Patwari, R. (2013). ROC Curves. Retrieved April 5, 2016, from https://www.youtube.com/watch?v=21Igj5Pr6u4&feature=youtu.be