Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data

Haowen You[1] and George Rumbe[2]
[1]*Department of Systems and Information Engineering*, University of Virginia, Charlottesville, Virginia.
[2]*Department of Systems Science and Industrial Engineering*, Binghamton University. Binghamton, New York.

Abstract* - *Accurate diagnostic detection of the cancerous cells in a patient is critical and may alter the subsequent treatment and increase the chances of survival rate. Machine learning techniques have been instrumental in disease detection and are currently being used in various classification problems due to their accurate prediction performance. Various techniques may provide different desired accuracies and it is therefore imperative to use the most suitable method which provides the best desired results. This research seeks to provide comparative analysis of Support Vector Machine, Bayesian classifier and other Artificial neural network classifiers (Backpropagation, linear programming, Learning vector quantization, and K nearest neighborhood) on the Wisconsin breast cancer classification problem.

Keywords: *Artificial Neural Networks, Classification, Breast cancer diagnosis*

I. INTRODUCTION

The development of automated diagnostics was instigated by the need to aid the physician in decision making. There application in healthcare has spanned from the electrocardiograms to ultrasounds etc. The traditional set-up for error detection and monitoring of disease progression heavily rest on the technicians within the healthcare. The increase in the number of patients within healthcare who require continuous assessment has led to the technical development of the automated systems. Transformations of the qualitative information to quantitative measures are at the forefront in solving classification problems. Breast cancer has been identified as the second largest cause of cancer deaths among women of age 40 and 55. The number of breast cancer diagnosis is estimated to be 1.2 million among women every year according to projections by the World Health Organization [4]. According to statistics by the American cancer society in 2001, about 40,200 deaths are caused by the breast cancers and 192,000 cases consist of women who are newly diagnosed [8]. Additional statistics as of 2006 estimated 214,460 new cancer diagnosis and total death at least 41,000 within the US [10]. Early detection and accurate diagnosis has been crucial in reducing the number of deaths which has increased the survival rate of those diagnosed with breast cancer [8].

The challenging effect of the identification of the cancerous cells in a patient is highly subjective and it is reliant on the physician expertise. This may lead to inaccurate predictions since the experiments are prone to human and visual error and may be affected by blurred mammogram visuals [11]. The aforementioned challenges necessitate the need for accurate tools for detection and classification of the breast cancer cells. There have been effective systems such as the machine decision support systems (MDSS) used in aiding breast cancerous cells detection [8]. Machine learning techniques have been instrumental in providing evidence in support of the accuracy of the classification of breast cancer patients. Once the breast cancer diagnosis has been performed the prognosis is subsequently determined to predict the future development and characteristics of the cancerous cells. Prognosis has been determined to be more complex due to the censoring of data [9].

Diagnosis is employed to significantly and accurately discern between malignant and benign cancerous patterns. Some of the conventional used approaches for breast cancer detection/diagnosis include mammography; surgical biopsy and fine needle aspirate [9]. The sensitivity results from the aforementioned approaches in accurately identifying the malignant lumps ranges as follows, mammography 68%-79%, fine needle aspirate 65%-98% and surgical biopsy about 100% [9]. The surgical biopsy despite being an effective approach has been determined as a costly procedure which induces negative psychological behavior on the patients [10]. Another effective method to diagnose breast masses is based on Fine Needle Aspiration biopsy, which is a technique to extract cell samples from lump and conduct vision observation on the cellular under microscope [1]. Diagnosis conclusion (benign and malignant) can be drawn according to the judgment of domain experts [2].

DOI: 10.9781/ijimai.2010.131

Currently, artificial intelligence techniques, which deal with the diagnosis as a pattern classification problem with the cellular nuclei shape information from cell slides images, have been introduced into this area, to improve the accuracy, consistency and efficiency of this diagnosis process.

A. Research Objective

The objective of this research is to provide a comparative study on the utilized potential classification tools (linear programming, back-propagation neural network, support vector machine and Bayesian network) on the problem by a benchmark dataset which consist of numeric cellular shape features extracted from preprocessed Fine Needle Aspiration biopsy image of cell slides.

B. Research scope

This research will first implement Support Vector Machine (SVM) and Bayesian network solution on the benchmark dataset. Then a comparison on this benchmark dataset between the former adopted techniques (linear programming and back-propagation neural network) and these two newly developed modeling approaches will be conducted. The measurement of this comparative study will be selected according to the proposed measures by the latest publication on this problem [4]. These will include classification accuracy, sensitivity, specificity, positive predictive value and negative predictive value. K-fold cross-validation [5] will also be used to evaluate the overall performance of each model built by aforementioned approaches. The organization for the rest of this research will be as follows, Section 2: provides detail information on the literature review, Section 3 introduces the strategies employed by the SVM and Bayesian network classifier, Section 4 discusses detail analysis on the results, the complexity of modeling process and the computation expenditure of these approaches, and Section 5 provides the summary and conclusion of the research.

II. Literature Review

The increase in the number of deaths determined within the healthcare systems has led to the development of medical diagnostic support systems to aid the medical personnel's in decision making process [10]. Various experts systems and machine learning algorithms have been utilized to provide supporting information based on the input knowledge. Some of the significant developments include 2D and 3D medical imaging, feature extraction, pattern analysis and classification have been used in providing solutions for edge detection and region growing among other problems [10]. According to Pena-Reyes and Sipper (1999) an effective diagnostic systems should be able to provide higher accuracy of disease identification as malignant or benign. In addition, the systems should also be able to determine with a degree of confidence indicating the accuracy of diagnosis with some levels. Another major important aspect is the systems interpretability which provides information on the steps followed resulting to the outcomes generated. The Artificial neural network on the other hand has been determined to be an effective tool in classification though the operations within the network structure are hidden.

Classification problem seems to have generated interests among researchers. The classification approach is used in data analysis and pattern recognition problems. This approach involves classifier modeling which is used as a function that associates a class to different attributes. The concept of association based on similarities or trained performance has been embedded in various approaches such as neural networks, decision trees, decision graphs and etc [14]. The methodology of the neural networks can be performed in two phases i.e. training and testing. The training phase involves feature extraction and computation utilizing the classification rules. On the other hand, testing data is used for performance evaluation on the accuracy of the classification process determined by the training data [10]. Breast cancer diagnosis and prognosis has instigated the research interest and has been explored utilizing various artificial neural networks such as Radial Basis Function, Multilayer perceptrons, Backpropagation, and Learning Vector Quantization network. Other methods which have been utilized to determine the breast cancer diagnosis includes Fuzzy systems and Evolutionary algorithms. The fuzzy systems are used to represents different degrees of the disease (malignant or benign) a patient suffers from; on the other hand, the evolutionary algorithms are used to perform search to determine the most suitable fuzzy systems [6].

Isotonic separation which is a linear programming technique is based on the underlying assumption of maintaining same consistency in diagnosis. For example the Breast cancer dataset (Wisconsin) a patients being diagnosed with malignant tumor based on certain characteristics of the cell structures, for other patients showing similar symptoms with more damage to the cells would end up receiving the same diagnosis [7] and Rank nearest neighbor technique (k-RNN) [11]. The k-RNN has been determined as technique used in approximating the densities based on the evaluations of the nearest neighbors [11]. The aforementioned technique has been applied in univariate and multivariate data in examining various classifications problems including breast cancer. In order for a patient to receive the appropriate breast cancer treatment, it is necessary that accurate classification of the cells be determined. This has lead researchers to combine and employ various machine learning techniques and select the one with the highest prediction accuracy [16]. The comparative analysis of the ANN ranges from two to

six networks or more being evaluated to determine the most appropriate technique. Integration of different ANN networks has lead to improve performance measures. The RBF properties when applied to tuning the SVM has been determined to provide higher prediction accuracy for breast cancer data [12].

III. METHODOLOGY

There have been numerous artificial neural network approaches used for examining the classification of breast cancer cells, some of these approaches are Bayesian classifier and SVM. This section provides descriptive discussions on the SVM and Bayesian classifier framework. In addition, it examines the strategies employed and some of the parameters that are used for effective classification of patterns.

A. 3.1 Support Vector Machine Stratagem

Support Vector Machine (SVM) was introduced by Vapnik and it is a technique based on the statistical learning theory and has been applied for solving classification and regression problems [15]. The objective of the SVM is to separate two classes by determining the linear classifier that maximizes the margin and it is referred to as the optimal separating hyperplane [15]. SVM has been employed in various classification problem and mostly current interest in breast cancer detection due its robustness. The regularization parameter and kernel function are the two major components that have to been determined before conducting training. Some of the significant researches employed using the SVM for breast cancer detection utilized heuristics SVM approaches such as the smooth SVM, the linear SVM and general non linear SVM [12]. The goal of SVM is to determine a suitable hyperplane with maximum margin which can be computed as an optimization problem [10].

B. 3.2 Bayesian Network Approach

Bayesian networks are characterized by the use of the probabilistic approach in problem solving and encompass the uncertainty of certain occurrences. Its origin is based on the probability distribution which can be depicted graphically. The Bayesian network classifier is composed of a set of variables related to each other by directed edges. The variables represent the data attributes, class and arcs, which when applied to the conditional probability table depicts their relationship in a visual format. The Bayesian network classifiers are also referred to as directed acyclic graphs that provide information on joint probability distribution on various random variables [14]. It has been determined that the Bayesian network classifier, the connecting arcs between different nodes provides an independence assumption that is associated with the different random variables. The independence assumption provides information on the probability distribution that is represented within the network. Generally, the probability distribution within the networks must initially be specified from the root nodes followed by the condition probabilities of the remaining non-root nodes based on the direct predecessor's combinations [13]. The conditional probabilities can only determine based on the fact that information on some of the nodes in the network have been identified.

The Bayesian network classifier uses the unsupervised learning algorithm, where the class target is unknown though we have the inputs (attributes) [14] and the classifier learning algorithm can be structured into two phases (i) Function for assessment of a certain network based on a data and (ii) an approach for examining space within the networks. There are various learning algorithm employed to the Bayesian network this includes AD (All dimensions) Trees, TAN (Tree Augmented Naïve Bayes) and K2. K2 has been used in breast cancer classification problems due to its fast convergence ability. Bayesian nets have been utilized in providing solutions to medical diagnosis, heuristic search and map learning problems among other challenges [13]. The Bayesian network is based on independence assumption between the nodes.

C. 3.3 Data Structure

The benchmark dataset in this research will obtained from the UCI Irvine machine learning repository http://archive.ics.uci.edu/ml/index.html. This dataset was originally created by Dr. Wolberg, Street and Mangasarian all from University of Wisconsin. Data items in the dataset are composed of ID number, the diagnosis which will either be classified as malignant (M) or benign (B) and numeric shape features of extract cellular nuclei such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, and symmetry and fractal dimension. The dataset was composed of a total of 569 observations with benign and malign cases being 357 and 212 observations respectively. Each of the dataset in the observation is composed of 30 variables and 10 of the featured variables are related to the aforementioned characteristics [3].

IV. RESULTS AND DATA ANALYSIS

This section provides discussion on the result and analysis for SVM, Bayesian, LVQ, KNN and BNT _Clustering. Furthermore, a comparative analysis of the aforementioned approaches is presented. The SVM and Bayesian network classifier approach were developed using MATLAB, and the 10 variables (see section 3.3) were experimented with within the classifiers.

A. 4.1 Support Vector Machine

Training for the SVM was conducted by varying a variety of C and gamma (ɣ) values based on 10 fold cross

validation. The ranges of C and ɣ were selected within the range of 2^{-15}- 2^{5} and 2^{3}-2^{15} respectively [18]. The two major SVM classifiers evaluated were C-SVM and Nu–SVM and the kernel functions that were used include polynomial, sigmoid and radial basis function. By examining the C-SVM employing the polynomial kernel function, the value of C=1 and ɣ= 2^{-3} showed 98.07% prediction accuracy which was the best from all other combinations. Figure 1 shows the surface plot for the varieties of C and ɣ. The initial values examined shows a flat surface which represents that the classification accuracy remained constant at 62.74% and progressively better predictions above 90% were determined.

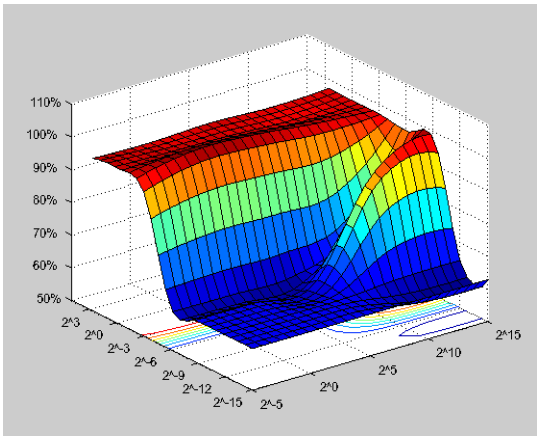


Figure 1: Surface plot for the C-SVM and Polynomial kernel function

The RBF kernel function was also examined and it showed better prediction accuracy as compared to the polynomial kernel function as shown in Figure 2. C=2^{15} and ɣ =2^{-15} showed a higher prediction accuracy of 98.24%. From Figure 2, the flat regions at the top indicate high accuracy prediction. The best prediction accuracy of 97.54% for the C-SVM using sigmoid kernel functions was determined between two regions (see Figure 3), i.e., when C=2^{10} and ɣ =2^{-6} and when C= 2^{10} and ɣ =$2^{-9.}$

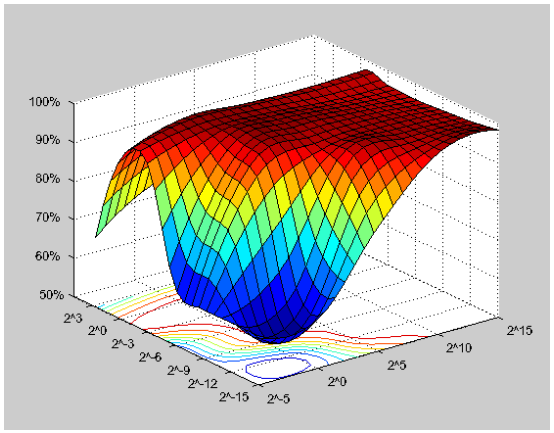


Figure 2: Surface plot for the C-SVM and RBF kernel functions

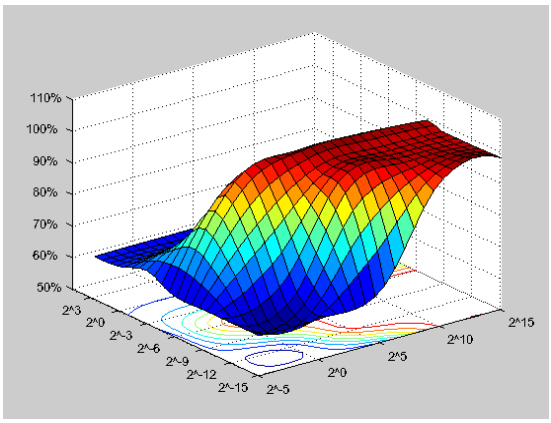


Figure 3: Surface plot for the C-SVM and sigmoid kernel function

Similar discussions were also presented using the Nu-SVM classifier with the polynomial, sigmoid and RBF kernel functions. The prediction accuracy of 92.79% was determined between the regions where C=2^{-1} and ɣ =2^{-6}, C=2^{15} and ɣ =2^{3} as shown Figure 4. The flat topmost regions which lies between the boundaries C=2^{15} and ɣ2^{-9} and 2^{3} and ɣ2^{3} showed a consistent prediction accuracy of more than 90%. Figure 5 shows the surface plot for the Nu-SVM and RBF kernel function which has a flat feature map with a small raised region due to high prediction accuracy above 90% obtained for the C and ɣ parameters. Prediction accuracy of 95.08 where C=2^{1} and ɣ =2^{3}, C=2^{5} and ɣ =2^{1}. A higher prediction accuracy of 93.67% using Nu-SVM and sigmoid kernel function was determined within the region where C=2^{15} and ɣ =2^{-15} as shown in Figure 6 below. Low prediction accuracy of less than 70% was obtained for values of C=2^{-5}-2^{15} and ɣ=1, and ɣ=8

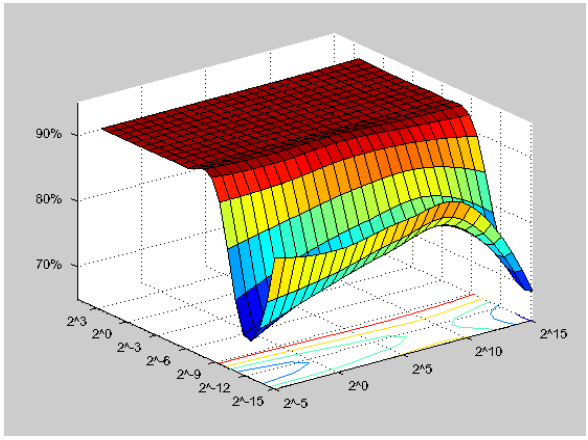


Figure 4: Surface plot for the Nu-SVM and polynomial kernel function

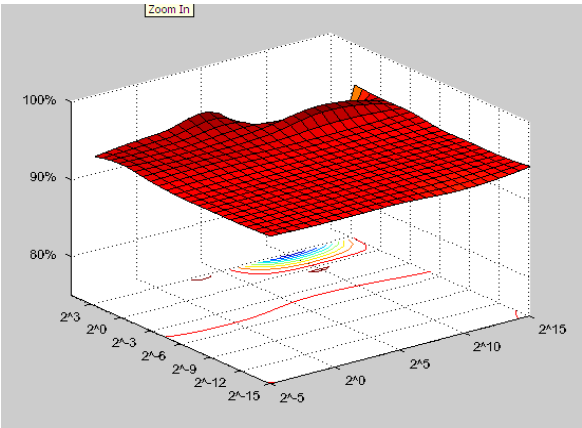


Figure 5: Surface plot for the Nu-SVM and RBF kernel function

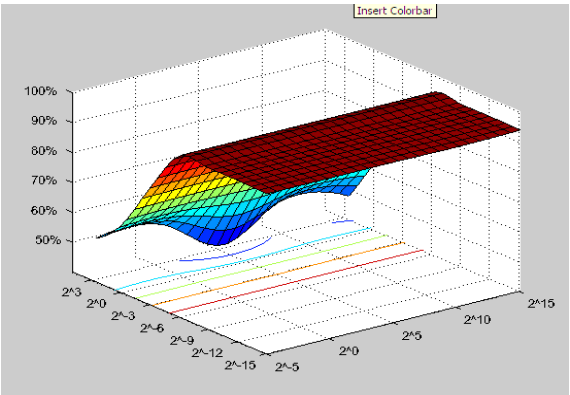


Figure 6: Surface plot for the Nu-SVM and sigmoid kernel function

B. Bayesian Network

The Bayesian network utilizes the Davies-Bouldin index during data preprocessing to change continuous data to discrete. In addition, Davies-Bouldin index assists in determining the appropriate cluster to be used in evaluating the network. The smaller the bouldin index indicates the most appropriate selection of the clusters. Figure 7, shows the Davies-Bouldin index which was utilized in this project.

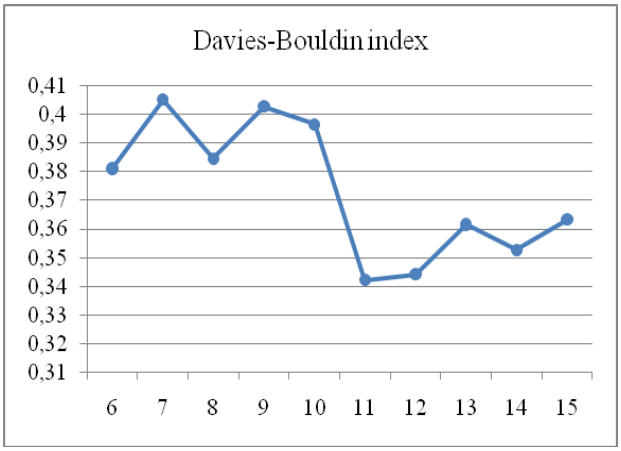
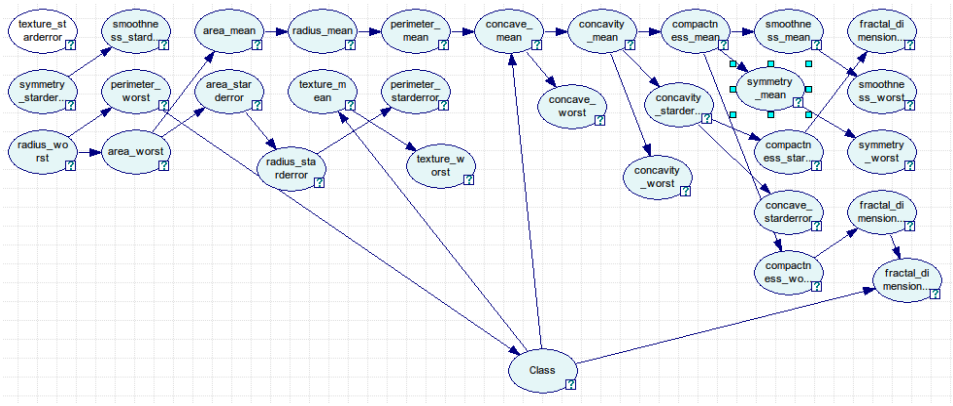


Figure 7: Davies-Bouldin index for different clusters on mean radius data item

The Bayesian network classifier was used for breast cancer classification. Three types of Bayesian network i.e., Naïve, K2 and Bdeu were examined to determine best network with higher prediction accuracy. The topologies for these different networks are shown in Figures 8 and 9. The topology for Naïve Bayes (see Figure 8) shows no learning takes place between input variables in the network. On the other hand, for K2 and Bdeu (see Figure 9) there is learning of relationship between the input variables. The experiments for each of the network were conducted by examining all the input features (All), mean and standard deviation (Mean+SE) and Mean. Results obtained from the network as illustrated in Figure 10 shows that Bdeu network with (All) had a higher prediction accuracy of 91.31%, followed by Naïve (All) at 89.55% and K2(Mean+SE) at 88.41%.

Figure 8: Naïve Bayesian network topology



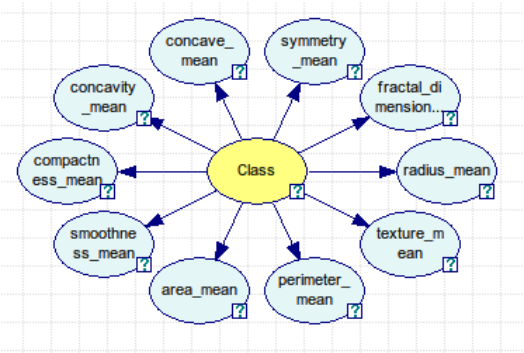


Figure 9: K2 and Bdeu network topology

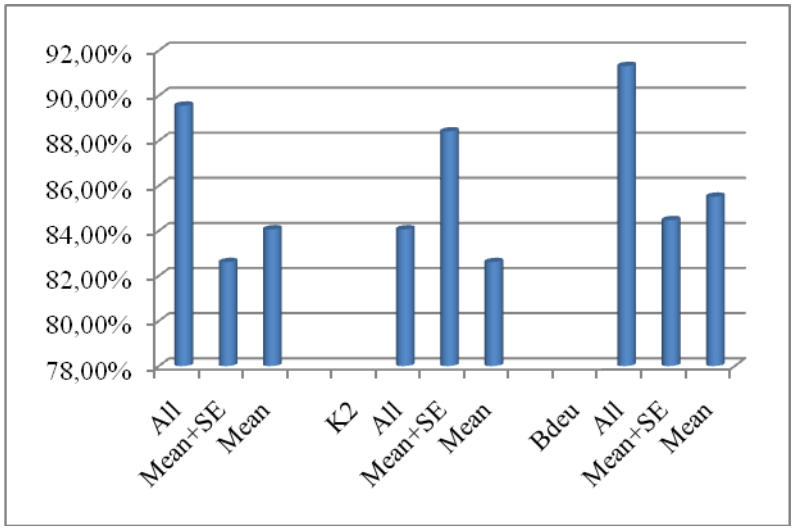


Figure 10: Bayesian network prediction results

C. Learning Vector Quantization

A combination of parameters of hidden neurons (5, 10, 15, 20, 25, and 30) and learning parameters (0.01, 0.1, 0.5, and 1) were varied against each other. The number of iterations for the network was set at 50. A higher prediction accuracy of 90.47% was determined with learning rate of 0.1 and 5 hidden neurons. Figure 11, shows the LVQ surface plot with low and high regions varying with the increase of learning rate and hidden neurons.

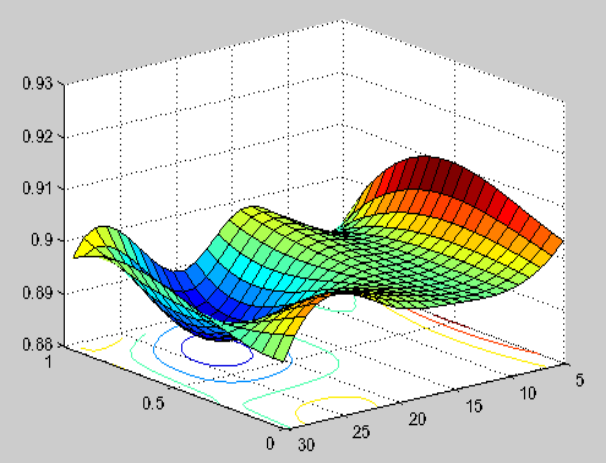


Figure 11: LVQ accuracy prediction surface plot

D. K-Nearest Neighborhood (KNN)

The KNN was evaluated using the Euclidean and Cityblock distance approach. The K (neighbors) evaluated ranged from 1 to 15. Figure 12 shows the results obtained and with a higher prediction accuracy being observed using both approaches. The Euclidean distance approach showed a prediction accuracy of 100% with K=5, 10 and 11, similarly to the Cityblock distance approach with K=13.

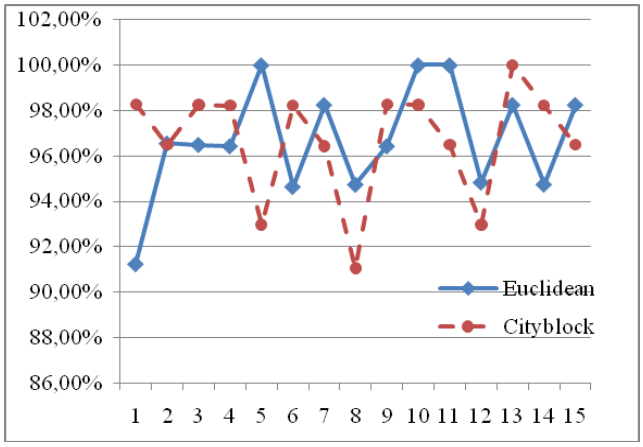


Figure 12: KNN prediction accuracy

E. Comparative Analysis

Table 1 shows a comparative analysis for the six different networks i.e., Support vector machines (SVM), Bayesian network (BNT), K nearest neighborhood (KNN), Learning vector quantization (LVQ), Linear programming (LP) and Backpropagation network (BPN). Based on the results in Table 1, the K nearest neighborhood had a higher prediction accuracy of 100%, followed by the SVM using the RBF kernel function with prediction accuracy of 98.24%. The K2 Bayesian network had poor prediction accuracy compared to all the networks evaluated. The results shows that machine learning techniques can provide accurate prediction and may enable proper classification of patient's condition and improve their quality of life. Although the Bayesian network classifier performance was poor compared to the SVM, the CPU time it took to produce the output results was low compared to other network. Table 2 shows the training and prediction time associated by each of the network observed in this project. The Bayesian network shows a low prediction time of 0.07seconnds.

Table 1: Comparative performance of breast cancer

Type	**SVM**		**BNT**			**KNN**		**LVQ**	**LP**	**BPN**
Kernel	C-SVM	Nu-SVM	Naïve	K2	Bdeu	Euclidean	CityBlock			
Polynomial	97.54%	92.79%	89.55 %	88.41 %	91.31%	100%	100%	91.04 %	97.50 %	95.33% [17]
RBF	98.24%	95.08%								
Sigmoid	97.72%	93.85%								

Table 2: Networks CPU time

Type of Network	**Training/Prediction Time (seconds)**
SVM	(2.11s)/0.94s
Bayesian Network Classifier	(4.27+1.51)s/0.07s
Learning Vector Quantization	67.18s/1.45
K Nearest neighbor	N/A/0.08s

V. Discussion and Conclusion

Early detection of breast cancer cells can be predicted accurately by the use of machine learning techniques. This may result in the decrease of health cost and may enhance time required for a patient to receive treatment. In this project the SVM and the Bayesian network have been discussed in providing diagnostic and prognosis assessment for breast cancer. The SVM has been determined to be more superior to Bayesian network since it provides higher prediction accuracy. By comparing the performance of both networks to other neural network approaches, the KNN has been examined to provide 100% classification. The prediction accuracy of the networks discussed in this project emphasizes the need of employing the machine learning techniques not only on the prediction of breast cancer data but on other medical conditions in which predictions of conditions are difficult to diagnose.

VI. References

[1] McMorran, J., Crowther., D.C., "Fine needle aspiration cytology (breast)", General Practice Notebook – a UK medical reference on the world wide web, Feb 2009.

[2] Olvi, L.M., Street, W.N., "Breast cancer diagnosis and prognosis via linear programming", Operations Research, Vol.43, No.4, 1995, pp. 570-577.

[3] UCI Irvine machine learning repository, "Breast Cancer Wisconsin (Diagnostic) Data Set", http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), Nov. 1995.

[4] Akay, M., "Support vector machines combined with feature selection for breast cancer diagnosis", Expert systems with applications, Vol.36, 2009, pp.3240-3247.

[5] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the fourteenth international joint conference on artificial intelligence, Vol.12, No.2, 1995, pp. 1137–1143.

[6] Pena-Reyes, C., and Sipper, M., A fuzzy approach to breast cancer diagnosis, Artificial intelligence medicine, Vol.17, 1999, pp.131-135.

[7] Ryu, Y., Chandrasekaran, R., and Jacob, V., Breast cancer prediction using the isotonic separation technique, European journal of operation research, Vol.181, 2007, pp.842-854.

[8] West, D., Mangiameli, P., rampal, R., West, V., "Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application", European journal of operation research, Vol.162, 2005, pp.532-551.

[9] Pantel, P., "Breast cancer diagnosis and prognosis", University of Manitoba (1998).

[10] Maglogiannis, I., and Zafiropoulos, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers", Application intelligence, Vol.30, 2009, pp.24-36.

[11] Bagui, S., Bagui, S., Pal, K., and Pal, N., "Breast cancer detection using rank nearest neighbor classification rules", Pattern recognition, 36, 2003, pp.25-34.

[12] Mu, T., and Nandi, A., "Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier", Journal of the Franklin Institute, Vol. 344, 2007, pp.285-311.

[13] Charniak, E., "Bayesian networks without tears," Artificial intelligence magazine, Vol.12, No.4, 1991, pp.50-63.

[14] Friedman. N., Geiger, D., and Goldszmidt, M., "Bayesian classifier", Machine learning, Vol. 29, 1997, pp.131-163.

[15] Gunn, S., "Support vector machines for classification and regression, Technical paper, 1998.

[16] Ubyeli. E., "Implementing automated diagnostic systems for breast cancer detection, Expert systems with application", Expert systems with applications, Vol.33, 2007, pp.1054-1062.

[17] Kim, Y., Jang, S., Cho, K., and Park, G., Performance comparison between Backpropagation, Neuro-Fuzzy Network, and SVM, Springer Berlin/Heidelberg, 2006.

[18] Hsu, C., Chang, C., and Lin, C., "A practical guide to support vector classification", Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003.http//www.csie.ntu.edu.tw/~cjlin/libsvm/.