**LITERATURE SURVEY**

**Machine learning for improved diagnosis and prognosis in healthcare**

Machine learning has gained tremendous interest in the last decade fueled by cheaper computing power and inexpensive memory – making it efficient to store, process and analyze growing volumes of data. Enhanced algorithms are being designed and applied on large datasets to help discover hidden insights and correlations amongst data elements not obvious to human. These insights help businesses take better decisions and optimize key indicators of interest. The growing popularity of machine learning also stems from the fact that learning algorithms are agnostic to the domain of application. Classification algorithms, for example, that could be applied to categorize faults in windmill blades can also be used for categorizing TV viewers in a survey. The actual value of machine learning however depends on the ability to adapt and apply these algorithms to solve specific real world problems. In this paper we discuss two such applications for interpreting medical data for automated analysis. Our first case study demonstrates the use of Bayesian Inference, a paradigm of machine learning, for diagnosing Alzheimer’s disease based on cognitive test results and demographic data. In the second case study we focus on automated classification of cell images to determine the advancement and severity of breast cancer using artificial neural networks. Although these research are still preliminary, they demonstrate the value of machine learning techniques in providing quick, efficient and automated data analysis. Machine learning offers hope with early diagnosis of diseases, help patients in making informed decisions on treatment options and can help in improving overall quality of their lives.

**Cancer statistics**

Each year, the American Cancer Society estimates the numbers of new cancer cases and deaths that will occur in the United States and compiles the most recent data on cancer incidence, mortality, and survival. Incidence data, available through 2014, were collected by the Surveillance, Epidemiology, and End Results Program; the National Program of Cancer Registries; and the North American Association of Central Cancer Registries. Mortality data, available through 2015, were collected by the National Center for Health Statistics. In 2018, 1,735,350 new cancer cases and 609,640 cancer deaths are projected to occur in the United States. Over the past decade of data, the cancer incidence rate (2005-2014) was stable in women and declined by approximately 2% annually in men, while the cancer death rate (2006-2015) declined by about 1.5% annually in both men and women. The combined cancer death rate dropped continuously from 1991 to 2015 by a total of 26%, translating to approximately 2,378,600 fewer cancer deaths than would have been expected if death rates had remained at their peak. Of the 10 leading causes of death, only cancer declined from 2014 to 2015. In 2015, the cancer death rate was 14% higher in non-Hispanic blacks (NHBs) than non-Hispanic whites (NHWs) overall (death rate ratio [DRR], 1.14; 95% confidence interval [95% CI], 1.13-1.15), but the racial disparity was much larger for individuals aged <65 years (DRR, 1.31; 95% CI, 1.29-1.32) compared with those aged ≥65 years (DRR, 1.07; 95% CI, 1.06-1.09) and varied substantially by state. For example, the cancer death rate was lower in NHBs than NHWs in Massachusetts for all ages and in New York for individuals aged ≥65 years, whereas for those aged <65 years, it was 3 times higher in NHBs in the District of Columbia (DRR, 2.89; 95% CI, 2.16-3.91) and about 50% higher in Wisconsin (DRR, 1.78; 95% CI, 1.56-2.02), Kansas (DRR, 1.51; 95% CI, 1.25-1.81), Louisiana (DRR, 1.49; 95% CI, 1.38-1.60), Illinois (DRR, 1.48; 95% CI, 1.39-1.57), and California (DRR, 1.45; 95% CI, 1.38-1.54). Larger racial inequalities in young and middle-aged adults probably partly reflect less access to high-quality health care. CA Cancer J Clin 2018;68:7-30.

**Using machine learning algorithms for breast cancer risk prediction and diagnosis.**

Breast cancer represents one of the diseases that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this paper, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed within a simulation environment and conducted in WEKA data mining tool.

**Comparative study of machine learning algorithms for breast cancer detection and diagnosis**

Breast cancer is one of the most widespread diseases among women in the UAE and worldwide. Correct and early diagnosis is an extremely important step in rehabilitation and treatment. However, it is not an easy one due to several uncertainties in detection using mammograms. Machine Learning (ML) techniques can be used to develop tools for physicians that can be used as an effective mechanism for early detection and diagnosis of breast cancer which will greatly enhance the survival rate of patients. This paper compares three of the most popular ML techniques commonly used for breast cancer detection and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN). The Wisconsin original breast cancer data set was used as a training set to evaluate and compare the performance of the three ML classifiers in terms of key parameters such as accuracy, recall, precision and area of ROC. The results obtained in this paper provide an overview of the state of art ML techniques for breast cancer detection.

**Using three machine learning techniques for predicting breast cancer recurrence.**

Abstract Objective: The number and size of medical databases are increasing rapidly but most of these data are not analyzed for finding the valuable and hidden knowledge. Advanced data mining techniques can be used to discover hidden patterns and relationships. Models developed from these techniques are useful for medical practitioners to make right decisions. The present research studied the application of data mining techniques to develop predictive models for breast cancer recurrence in patients who were followed-up for two years. Method: The patients were registered in the Iranian Center for Breast Cancer (ICBC) program from 1997 to 2008. The dataset contained 1189 records, 22 predictor variables, and one outcome variable. We implemented machine learning techniques, i.e., Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN) to develop the predictive models. The main goal of this paper is to compare the performance of these three well-known algorithms on our data through sensitivity, specificity, and accuracy. Results and Conclusion: Our analysis shows that accuracy of DT, ANN and SVM are 0.936, 0.947 and 0.957 respectively. The SVM classification model predicts breast cancer recurrence with least error rate and highest accuracy. The predicted accuracy of the DT model is the lowest of all. The results are achieved using 10-fold cross-validation for measuring the unbiased prediction accuracy of each model.

**Machine learning based performance development for diagnosis of breast cancer**

Breast cancer is prevalent among women and develops from breast tissue. Early diagnosis and accurate treatment is vital to increase the rate of survival. Identification of genetic factors with microarray technology can make significant contributions to diagnosis and treatment process. In this study, several machine learning algorithms are used for Diagnosis of Breast Cancer and their classification performances are compared with each other. In addition, the active genes in breast cancer are identified by attribute selection methods and the conducted study show success rate 90,72 % with 139 feature.

**A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis.**

Breast cancer is becoming a leading cause of death among women in the whole world, meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients. Expert systems and machine learning techniques are gaining popularity in this field because of the effective classification and high diagnostic capability. In this paper, a rough set (RS) based supporting vector machine classifier (RS\_SVM) is proposed for breast cancer diagnosis. In the proposed method (RS\_SVM), RS reduction algorithm is employed as a feature selection tool to remove the redundant features and further improve the diagnostic accuracy by SVM. The effectiveness of the RS\_SVM is examined on Wisconsin Breast Cancer Dataset (WBCD) using classification accuracy, sensitivity, specificity, confusion matrix and receiver operating characteristic (ROC) curves. Experimental results demonstrate the proposed RS\_SVM can not only achieve very high classification accuracy but also detect a combination of five informative features, which can give an important clue to the physicians for breast diagnosis.

**SVM and SVM ensembles in breast cancer prediction**

Breast cancer is an all too common disease in women, making how to effectively predict it an active research problem. A number of statistical and machine learning techniques have been employed to develop various breast cancer prediction models. Among them, support vector machines (SVM) have been shown to outperform many related techniques. To construct the SVM classifier, it is first necessary to decide the kernel function, and different kernel functions can result in different prediction performance. However, there have been very few studies focused on examining the prediction performances of SVM based on different kernel functions. Moreover, it is unknown whether SVM classifier ensembles which have been proposed to improve the performance of single classifiers can outperform single SVM classifiers in terms of breast cancer prediction. Therefore, the aim of this paper is to fully assess the prediction performance of SVM and SVM ensembles over small and large scale breast cancer datasets. The classification accuracy, ROC, F-measure, and computational times of training SVM and SVM ensembles are compared. The experimental results show that linear kernel based SVM ensembles based on the bagging method and RBF kernel based SVM ensembles with the boosting method can be the better choices for a small scale dataset, where feature selection should be performed in the data pre-processing stage. For a large scale dataset, RBF kernel based SVM ensembles based on boosting perform better than the other classifiers.