

Disease Prediction in Data Mining Technique using Machine Learning

¹ CHANDRA BHUSHANA RAO KILLI, ² K. KANTHARAJU

¹ Associate Professor,

Department of Computer Science and Engineering,

GDMM College of Engineering and Technology, Nandigama, Krishna Dist,
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India.
kchbhushan.mtech@gmail.com

² Assistant Professor and HOD

Department of Computer Science and Engineering,

GDMM College of Engineering and Technology, Nandigama, Krishna Dist,
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India.
kantaraju8@gmail.com

Abstract- Various Cancer is now a day's one of the main diseases which has widely affected among the peoples in our globe. A molecular pathologist selects a list of genetic variations of interest that he/she wants to analyze. The molecular pathologist searches for evidence in the medical literature that somehow is relevant to the genetic variations of various interest finally this molecular pathologist spends a huge amount of time detecting the evidence which is related to each of the variations to classify them. The ultimate goal is to replace proposed model by a machine learning model.

The molecular pathologist will still have to decide which variations area of interest, and also collect the relevant evidence. In this paper, we apply machine learning methods especially logistic regression (which is more accurate) on the datasets to determine and examine whether there are any signs or possibilities of cancer and if the person is examined as cancerous then the stage of cancer is also determined. Cancer disease is classified into four types named type 1, type 2, type 3 and type 4. Id, Gene, variation, and class are the fields used.

Keywords: Machine Learning, Breast cancer, Prostate cancer, SVM, Cancer

1. INTRODUCTION

Machine learning contains a set of methods, which allow a machine to learn meaningful patterns from data directly with minimal human interaction. The strength of a machine learning technique is, in part, dependent on human knowledge. Such knowledge can help a machine to learn more efficiently through techniques like appropriate feature selection, transfer learning, and multitask learning. Through this symbiosis, machine learning has been successfully applied in many applications and achieves state-of-the-art performance. More

recently, machine-learning techniques have been applied to the field of medical imaging [1].

With fast improving computational power and the availability of enormous amounts of data, deep learning has become the default machine-learning technique that is utilized since it can learn much more sophisticated patterns than conventional machine-learning techniques. Unlike conventional machine-learning techniques, deep learning methods greatly simplify the feature engineering process and some have even been applied to raw data directly. This is especially important for the field of medical imaging analysis since it can take years of training to obtain adequate domain expertise for appropriate feature determination.

Hence, this allows more researchers to exploit new ideas easier and faster. Among all deep learning methods, convolution neural networks (CNNs) are of special interest. By exploiting local connectivity patterns efficiently with shared weights, CNN, such as those utilized in the Image Net competition, has quickly become a state-of-the-art method for image processing. Naturally, there are many recent works trying to apply CNN on medical image analysis.

With methods like the rectified linear unit and deep residual learning alleviating issues such as the vanishing gradient problem, deeper models can be trained more efficiently and hence pushing deep learning to another level [2].

However, there are still many remaining challenges, e.g., inconsistencies in data formats and lack of reliable training data, which need to be addressed. An active research topic is how to optimize the transfer of human knowledge to a machine-learning model. % is special issue focuses on applying machine-learning techniques to medical imaging data and covers topics from traditional machine-learning techniques, e.g., principle component analysis and support vector machine, to more recent ones, such as CNN.

Transfer learning, which is used to address the issue of lacking sufficient medical image data for training, is also discussed. With the successful application of these techniques, papers in this special issue show progress on many fronts, such as the diagnosis of Alzheimer's disease and liver tumor segmentation. We hope that the readers will find these topics interesting [3-6].

2. LITERATURE SURVEY

In recent years, few works have been reported in the literature for the design and development for the analysis and detection of the various types of cancer cells present. Various reports have been created and much hard work has already been performed for implementing the cancer detection model. In 2002, Isabelle Guyon Jason Weston Stephen Barn hill says DNA micro-arrays now permit scientists to screen thousands of genes simultaneously and decide active hyperactive cancerous tissue.

Because new micro-array devices generate bewildering amounts of raw data, new analytical methods must be developed to sort out whether cancer tissues have distinctive signatures of gene expression over normal tissues or other types of cancer tissues. In contrast with the baseline method, their method eliminates gene redundancy automatically and yields better and more compact gene subsets.

In patients with leukemia (lack of white blood cells, (WBC)) our method discovered 2 genes that yield zero leave-one-out error, while 64 genes are

necessary for the baseline method to get the best result (one leave-one-out error). Using 4 genes from the colon cancer database, this method is 98% accurate, while the baseline method is only 86% accurate [4].

In 2003, Ain Choon tan and David Gilbert In recent years, computational diagnostic tools and artificial intelligence techniques provide automate procedures for objective judgments by making use of quantitative measures and machine learning techniques. In this paper, they proposed a Support Vector Machines (SVMs) based classifier in comparison with Bayesian classifiers and Artificial Neural Networks for the prognosis and diagnosis of breast cancer disease. This paper was quite helpful for the classifiers provided with the implementation details and the corresponding results.

Many comparative studies have been implemented regarding both the prognosis and diagnosis problem narrating the superiority of the proposed SVM algorithm in terms of accuracy. In 2007, Ilias Maglogiannis·Elias Zafiroopoulos· Ionize Anagnostopoulos says In few years, computational diagnostic tools and artificial intelligence techniques provide automated procedures for objective judgments by making use of quantitative measures and machine learning techniques [5].

In this paper, they proposed a Support Vector Machines (SVMs) based classifier in comparison with Bayesian classifiers and Artificial Neural Networks for the problem demonstrating the superiority of the proposed SVM algorithm in terms of accuracy. In 2009, Jagpreet Chhatwal^{1,2,3} Oguzhan Alagoz² Mary J. Lindstrom says descriptors of the National Mammography Database using the algorithm logistic regression that can aid in decision making for the early detection of breast cancer.

We created two logistic regression models based on the mammography features and demographic data for 62,219 consecutive mammography records from 48,744 studies in 18,270 patients had been reported

using the Breast Imaging Reporting and Data System (BI-RADS) lexicon. State cancer registry outcomes were matched with the data served as the reference standard [7].

Then it was found that the probability of cancer is found in both the models. Thus the variables of model 1 and the radiologist BI-RADS assessment were togetherly used for building Model2. We have used the algorithm 10-fold cross-validation to train and test the model and to calculate the area under the receiver operating characteristic curves (Az) to measure the performance. Both models were compared with the radiologists' BI-RADS assessments and accuracy was achieved.

In 2015[4], Rajesh Kumar, Rajeev Srivastava, and Subodh Srivastava say a framework for automated detection and classification of cancer by microscopic biopsy images using clinically significant and biologically interpretable features is proposed and examined. The various stages involved in the methodology include enhancement of microscopic images, segmentation of background cells, features extraction, and finally the classification [8].

An appropriate and efficient method is employed in each of the design cells, features extraction, and finally the classification. An appropriate and efficient method is employed in each of the steps of the proposed framework after making a comparative analysis of the commonly used method in each category

3. BACKGROUND

Detection and analysis of the disease Cancer have always been a major issue for the pathologists and the medical practitioners for treatment planning. Tumor grade is a tumor based on how abnormal the tumor cells and the tumor tissue look under a microscope. It provides us an indication of the growth of the tumor. If the cells of the tumor and the organization of the tumor's tissue resemble those of normal cells and tissue, then the tumor is called as "well differentiated". As compared with the

undifferentiated" or "poorly differentiated," these tumors tend to have less capacity to grow and spread. Based on differences in microscopic appearance, doctors assign a numerical "grade" to most of the cancers.

The factors used to determine tumor grade grow rapidly and spread faster than tumors with a lower grade. If a grading system for a tumor type is not specified, then the following system is generally used GX: Grade cannot be assessed (undetermined grade)

G1: Well differentiated (low grade)

G2: Moderately differentiated
(intermediate grade)

G3: Poorly differentiated (high grade)

G4: Undifferentiated (high grade)

Breast and prostate cancers are the most common types of cancer as they have their grading systems. Tumors present Different cancers are classified concerning the grading system.

Breast cancer:

Doctors mostly use the Nottingham grading system (also known as Elston-Ellis modification of the Scarff-Bloom-Richardson grading system) for breast cancer. This system grades breast tumors based on the following factors[1,3,4]. Tubule formation, that is, how much of the tumor tissue has normal breast (milk) duct structures Nuclear grade: It is an evaluation of the size and shape of the nucleus in the tumor cells Mitotic rate: The number of dividing cells present, which results in the speed of growing and dividing the cells.

Each of the categories gets a score between 1 and 3; the most abnormal. The scores for the three categories are added, having a total score of 3 to 9. Three grades are possible: Total score = 3–5: G1 denotes low grade or well differentiated. Total score = 6–7: G2 denotes intermediate-grade or moderately differentiated Total score = 8–9: G3 denotes high grade or poorly differentiated [7].

Issues and Challenges:

- In the author discussed and resolved the algorithms, issues and techniques for the problem of breast cancer prediction. This analysis does not include records with missing data so in future this work is enhanced by including the missing data.
- In by analyzing the artificial neural network, trained on a number of clinic pathological variables of patients with breast cancer, predicted survival with high accuracy. The author concluded that the consistent accuracy over time and the good predictive performance of a network trained without information on nodal status. It shows that neural networks are valuable tools in cancer survival prediction. In future the study should concentrate on collecting data from a more recent time period and find new potential prognostic factors to be included in a neural network model.
- In the survey, the study is based on multiple prediction models for breast cancer survivability using large datasets along with 10 fold cross validation method. It provides a relative prediction ability of different data mining methods. In future this work is extended by collecting real dataset in the clinical laboratory.

Prostate cancer:

Prostate cancer is graded by the Gleason scoring system. The Gleason score is based on biopsy samples that are taken from the prostate. The pathologist checks the samples to see how much similar the tumor tissue resembles normal prostate tissue. Both the primary and the secondary pattern of tissue organization are identified. The primary pattern represents and signifies the most common tissue pattern that is seen in the tumor, and the secondary pattern represents the next pattern. Each pattern is given a grade from 1 to 5, with 1 resembling the most like prostate tissue and 5

resembling the most abnormal. Gleason score is obtained by adding the two grades.

The American Joint Committee on Cancer advised to group Gleason scores into the following categories:

- Gleason X: Gleason score cannot be determined
- Gleason 2–6: The tumor tissue is well-differentiated
- Gleason 7: The tumor tissue is moderately differentiated
- Gleason 8–10: The tumor tissue is poorly differentiated.

Doctors often use tumor grade and other factors, such as cancer stage and a patient's age and general health, to develop a treatment plan and to determine a patient's prognosis.

Generally, a lower grade indicates a better prognosis. Higher-Grade cancer may grow and spread more quickly and require immediate aggressive treatment. The importance of tumor grade in planning treatment and determining a patient's prognosis is more for certain types of cancer, such as soft tissue sarcoma, primary brain tumors, and breast and prostate cancer. Patients should consult with their doctor for more information about tumor grades and moreover their relation with their treatment and prognosis [9].

Heart Disease Prediction

Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis for widely distributed in raw medical data which is heterogeneous in nature and voluminous.

These data should be collected in an organized form. This collected data can be integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data.

From the analysis of World Health Organization, they estimated 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths occur in United States and other developed countries due to cardio vascular diseases [10].

On the above discussion, it is regarded as the primary reason behind deaths in adults. Heart disease kills one person every 34 seconds in the United States. The following paper reviewed about predicting of heart disease using data mining technique [6].

Support Vector Machine (SVM)

In 1979 support vector machine was firstly discovered by vapnik[26]. Support vector machine is learning machine that used as a tool for the function approximation, classification of data, etc, because of its generalization capacity it constitute success in multiple applications. SVM has advantage of the selection of model automatically which means that the locations of basic functions and optimal number were automatically generated during training. Working of SVM was mostly based on kernels.

SVM can use for the pattern classification which has MLP and RBF networks. The SVM is the leading technology that including maximal classification algorithms that are embedded in structural learning theory. SVM method is using for the classification of non-linear and linear data. SVM converts original training dataset into high dimension by use of non-linear mapping. With this newly generated dimension SVM researches for linear optimal severing hyperplane.

Data from the two classes are separated by the hyper plane by using a convenient nonlinear mapping for an adequately high dimension. By using support vectors & margins SVM determine this hyper plane. SVM implements classification function by magnifying the margin that classifies both the class although minimizing classification errors. SVM can also apply to the different

optimization tasks like regression; classic complication is that of data classification.

Algorithm:

- Step1: Load all the necessary libraries -numpy, pandas, matplotlib
- Step 2: Load all the datasets in the machine.
- Step 3: Input and output Data
- Step 4: Perform normalization
- Step 5: Split the dataset into the train set and test set.
- Step 6: calculate z value.
- Step 7: Perform prediction based on the given dataset and check the accuracy of the technique.

4. CONCLUSION

The contribution of our study is to prove the feasibility and effectiveness of machine learning for the detection and analysis of different types of cancer. In this paper, we employed simple classification techniques that are mainly KNN, SVM, Logistic Regression, Linear regression.

While performing this machine learning model for cancer detection, we got the most accurate results in logistic regression. Thus in the future, we will perform and implement fewer ideas for easier conductance for this Machine learning model.

References

- [1]. Carloz Ordonez, "Association Rule Discovery with Train and Test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006, pp 334-343.
- [2]. M. ANBARASI, E. ANUPRIYA, N.CH.S.N.IYENGAR, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370- 5376
- [3]. G. Parthiban, A. Rajesh, S.K.Srivatsa "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method"
- [4]. Choi J.P., Han T.H. and Park R.W., "A Hybrid Bayesian Network Model for Predicting Breast

- Cancer Prognosis”, J Korean Soc Med Inform, 2009, pp. 49-57
- [5]. Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining, 2006.
- [6]. Lundin M., Lundin J., Burke B.H., Toikkanen S., Pylkkänen L. and Joensuu H. , “Artificial Neural Networks Applied to Survival Prediction in Breast Cancer”, Oncology International Journal for Cancer Research and Treatment, vol. 57, 1999.
- [7]. Delen Dursun , Walker Glenn and Kadam Amit , “Predicting breast cancer survivability: a comparison of three data mining methods,” Artificial Intelligence in Medicine ,vol. 34, pp. 113-127 , June 2005.
- [8]. Ruben D. Canlas Jr., “DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES”, August 2009
- [9]. Michael Feld, Dr. Michael Kipp, Dr. Alassane Ndiaye and Dr. Dominik Heckmann “Weka: Practical machine learning tools and techniques with Java implementations”
- [10]. K.P Soman, Shyam Diwakar, V.Vijay “Insight into Data mining theory and practice”