OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

A SURVEY ON IMPROVING THE PERFORMANCE OF CANCER AND DIABETES DETECTION BY USING NOVEL TECHNIQUE OF MACHINE LEARNING

Samrudhi Kaware[1], Dr. V. S. Wadne[2]

PG Student, Department of Computer Engineering, ICOER Wagholi, Pune[1]
Associate Professor, Department of Computer Engineering, ICOER Wagholi, Pune[2]

samrudhikaware17@gmail.com[1], vinods1111@gmail.com[2]

***ABSTRACT:* The system allows the user to make a use of algorithms to predict the risk of diabetes mellitus in human body. The various classification models such as Decision Tree, artificial Neural Networks, Logistic Regression, Association rules and Naive Bayes are used in this system. Then the Random Forest technique is used to find the accuracy of each model in the system. The dataset used is the Pima Indians Diabetes Data Set, which has the information of patients, some of them have developing diabetes therefore, this project is aimed to create a mobile application for predicting a person's class whether present in of the diabetes and cancer risk or not.**

Keywords: *Decision tree, ANN, Health-Care*

I INTRODUCTION

Cancer and diabetes are two destructive diseases in our society. Every year numerous people die out of cancer. The Agency for Healthcare Research and Quality (AHRQ) says that medical cost for cancer in the year 2011 in the United States was 88.7 billion dollars [1]. And out of various types of cancer, breast cancer has been one of the significant types over the past years [2]. Sometimes, breast cancer is detected at a stage when chances of survival are very low. Computer science can play some role to detect vulnerability of a cancer patient with medical data with the help of machine learning. By manipulating medical data, having attributes of cancer cell, a system can predict if the cancer is benign or malignant [3]. If the cancer is in a benign stage, then taking appropriate measures can help the patient survive and can even heal them completely in some cases. Cancer and diabetes is another disease that kills people slowly. Cancer and diabetes has become prevalent almost all over the world [4]. However, according to a study of Asian. Diabetic Prevention Organization, 60 percent of the whole world's diabetic population is from Asia [5]. So, Asian people are at high risk. Existence of cancer and diabetes in a patient can be predicted by machine learning. So, in this research, cancer and diabetes is predicted as binary values like 1 or 0 meaning "YES" or "NO". The data set that is used for cancer and diabetes includes attributes of the patients' feature that might lead to the existence of cancer and diabetes. Machine learns the attributes and then predicts in "YES" or "NO". The main objective of the research is to predict Cancer and Cancer and diabetes. For cancer it will predict the stage as "Malignant" or "Benign" and for cancer and diabetes it will predict as "YES" or "NO". The prediction is based on some of the state of the art machine learning algorithms. The project has another objective as to optimize the performances of these well-established machine learning algorithms. Some experiments will be performed to see if the algorithms can perform better on a different setup. Performance comparison is checked across different classifiers to understand how they behave with the same data set and how much time does each one take to build a classification model. One challenging prospect of this project is to achieve some techniques to apply curriculum learning [6] on the data set.

II RELATED WORK

Techniques used for Cancer and Diabetes Prediction

Abundant literature has been dedicated to the classifiers of data mining and tremendous progress has been made ranging from efficient and scalable algorithm for different datasets. These chapters provide a brief overview of the current status of diabetes prediction and discuss a few promising research directions. We believe that mining research has substantially broadened the scope of data analysis and will have deep impact on mining methodologies and applications in the future. However, there are still some challenging research issues that need to be solved in searching using concept of various classifiers of data mining in the research of diabetes prediction.

K-means Algorithm and Logistic Regression

A novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). Based on a series of pre-processing procedures, the model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were utilized to compare results with the results from other researchers.

Temporal Abstraction with Data Mining

Hemodialysis patients might suffer from unhealthy care behaviors or long-term dialysis treatments and need to be hospitalized. If the hospitalization rate of a hemodialysis center is high, its service quality will be low. Therefore, decreasing hospitalization rate is a crucial problem for health care centers. This study combines temporal abstraction with data mining techniques for analyzing dialysis patients' biochemical data to develop a decision support system. The mined temporal patterns are helpful for clinicians to predict hospitalization of hemodialysis patients and to suggest immediate treatments to avoid hospitalization.

Electromagnetism-like Mechanism

The use of artificial intelligence based data mining techniques for massive medical data classification and diagnosis has gained its popularity, whereas the effectiveness and efficiency by feature selection is worthy to further investigate. They presented a novel method for feature selection with the use of opposite sign test (OST) as a local search for the electromagnetism-like mechanism (EM) algorithm, denoted as improved electromagnetism- like mechanism (IEM) algorithm. Nearest neighbor algorithm is served as a classifier for the wrapper method. The proposed IEM algorithm is compared with nine popular feature selection and classification methods. Forty-six datasets from the UCI repository and eight gene expression micro array datasets are collected for comprehensive evaluation. Non-parametric statistical tests are conducted to justify the performance of the methods in terms of classification accuracy and Kappa index. The results confirm that the proposed IEM method is superior to the common state-of- art methods.

Pre-Diabetes Detection by Risk Factors

Purpose of this was to compare the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes or pre-diabetes using common risk factors. A standard questionnaire was administered to obtain information on demographic characteristics, family diabetes history, anthropomorphic measurements and lifestyle risk factors. Then He developed three predictive models using 12 input variables and one output variable from the questionnaire information we evaluated the three models in terms of their accuracy, sensitivity and specificity. The logistic regression model achieved classification accuracy.

III PROPOSED SYSTEM

The process starts with data manipulation. Next, four models will be investigated for finding a prediction model. Then, accuracy of each model will be calculated and compared for seeking the best model. Detecting diseases like cancer and diabetes might be helpful for the patients as well as the doctors. From the doctors' perspective, they can help the patients to identify their next step by identifying the vulnerability of cancer or prevalence of diabetes in a patient. The study ends up with creating a web application.

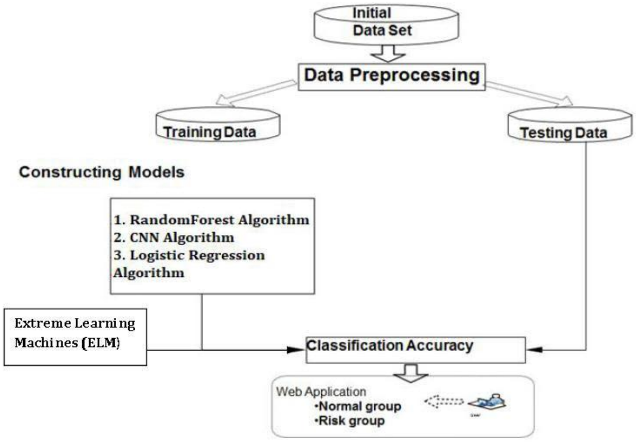

Figure 1: System Architecture

IV CONCLUSION AND FUTURE WORK

This system aimed to establish an appropriate prediction model for the high-risk T2DM group. Based on a number of researchers' experiences, authors proposed a novel model, which consists of double-level algorithms, i.e., the improved K-means and logistic regression algorithms. In order to make a valid comparison with others' results, it was necessary to conduct this model using the WEKA toolkit and use the same Pima Indian Diabetes Dataset. Proper filters were utilized to improve the validity and rationality of the dataset. The proposed model that consisted of both cluster and class method ensured the enhancement of prediction accuracy. Another realistic dataset provided by Dr. Schorling was used to test and verify the model. This proposed model has proven to be appropriate for predicting T2DM. One of proposed model's benefits is that it avoids deleting overmuch original data. It ensures the high quality of experimental data. The other benefit is that model can apply in the Pima Indian Diabetes Dataset as well as other various datasets. While the limitation is that it consumes more time during the part of pre-processing. We described that some papers focus on improving K-means by optimizing the initialized procedure of cluster center. But this improved model is based on the purpose of predicting DM2 and matches up with the logistic regression algorithm. It assures less time consuming and maximum retention of original data. Although the improved model is not so complicated, it attained well effect according to plenty of experiments. The main problems solved are improving accuracy of prediction model and making the model to adapt to different datasets. In this seminar, it was concluded that proposed model showing higher prediction accuracy than other researchers' experimental results. And the improved Kmeans algorithm we proposed contributed a lot to the prediction model. Moreover, there are two more dataset applied in proposed model and all of them obtained well effect.

REFERENCES

[1] Mustakim Al Helal, Atiqul Islam Chowdhury , Ashraful Islam , Eshtiak Ahmed , Md. Swakshar Mahmud, Sabrina Hossain "An Optimization Approach to Improve Classification Performance in Cancer and Diabetes Prediction" International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[2] American Cancer Society, "Cancer Facts & Figures 2015," Cancer Facts Fig. 2015, pp. 1–9, 2015.

[3] BCSC, "Types of Breast Cancer," Breast Cancer Society of Canada, 2014. [Online]. Available: http://www.bcsc.ca/p/41/l/506/t/Breast-Cancer-Society-of-Canada ---Types-of-Breast-Cancer..

[4] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification" Expert Syst. Appl., vol. 41, no. 5, pp. 2239–2249, 2014.

[5] Eftychios A. Pnevmatikakis, Petros Maragos "An Inpainting System For Automatic Image Structure-Texture Restoration With Text Removal", IEEE trans. 978-1-4244-1764, 2008

[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods", Computer. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014.

[7] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes.," BMC Med. Inform. Decis. Mak., vol.10, p. 16, 2010.