BOSOM DISEASE DETECTION USING ML TECHNIQUES

Dr. B. Narendra Kumar, Professor , Sridevi Women's Engineering College, Hyderabad, E-mail: swecnarendra@gmail.com
Badam Sneha, B.Tech, Dept of Information Technology, Sridevi Women's Engineering College, Hyderabad, E-mail: snehabadam333@gmail.com
Kunisetti Tabu Sri, B.Tech, Dept of Information Technology, Sridevi Women's Engineering College, Hyderabad, E-mail: tabusrikunisetti@gmail.com
Tummanapalli Roshini, B.Tech, Dept of Information Technology, Sridevi Women's Engineering College, Hyderabad, E-mail: roshinitummanapalli@gmail.com
Lingatla Sathvika, B.Tech, Dept of Information Technology, Sridevi Women's Engineering College, Hyderabad, , E-mail: lingatlasathvika@gmail.com

ABSTRACT: Cancer is the common problem for all people in the world with all types. Particularly, Breast Cancer is the most frequent disease as a cancer type for women. Therefore, any development for diagnosis and prediction of cancer disease is capital important for a healthy life. Machine learning techniques can make a huge contribute on the process of early diagnosis and prediction of cancer. In this paper, two of the most popular machine learning techniques have been used for classification of Wisconsin Breast Cancer (Original) dataset and the classification performance of these techniques have been compared with each other using the values of accuracy, precision, recall and ROC Area. The best performance has been obtained by Support Vector Machine technique with the highest accuracy.
Keywords: machine learning; breast cancer; classification; early diagnosis.

1. INTRODUCTION

Cancer is the second reason of human death all over the world and accounts for roughly 9.6 million deaths in 2018. Globally, for 1 human death in 6 can be said that is caused by cancer. Almost 70 percent of the deaths from cancer disease happen in countries that have low and middle income [1]. The most common cancer type among women are breast, lung and colorectal, which totally symbolize half of the all cancer cases. Also, breast cancer is responsible for the thirty percent of all new cancer diagnoses in women [2]. Machine learning (ML) methods ensure analyzing the data and extracting key characteristics of relationships and information from dataset. Also, it creates a computational model for best description of the data. Especially, according to in researches about cancer disease, it can be said that ML techniques can be handled on early detection and prognosis of cancer [3]. Bazazeh and Shubair have investigated the comparative study of machine learning techniques as Support Vector Machine (SVM), Random Forest (RF) and Bayesian Network (BN) for detection and diagnosis of breast cancer. The Original Wisconsin Breast Cancer was used as a dataset and Weka software was used as a Machine Learning tool.

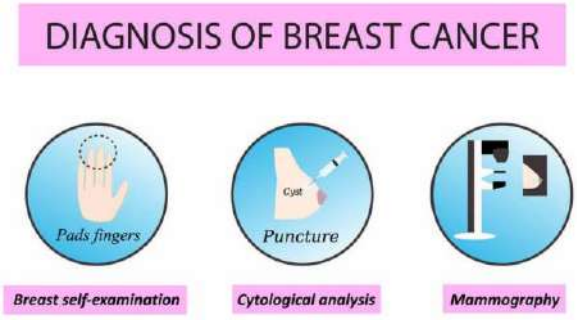

Fig.1 Diagnosis of breast cancer

The key performance parameters of machine learning classifiers have been compared according to accuracy, recall, precision and ROC area. They have suggested that BN has the best performance according to recall and precision values and RF technique has optimum performance in term of ROC area [4]. Ahmad et al. have

exercised machine learning algorithms for predicting the rate of two years recurrence of breast cancer disease. The dataset has been obtained from Iranian Center of Breast Cancer (ICBC) program, collected the time period of 1997-2008 years. The dataset is consisted of population characteristics and 22 input variables also the cases have been collected from 1189 women of diagnosed breast cancer. Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) have been applied and SVM has been showed the best performance with highest accuracy and least error rate [5].

In this work, SVM and ANN two of the most popular machine learning techniques ate applied on WBC dataset and the result of applied ML techniques are compared according to performance metrics.Support Vector Machine (SVM) is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Artificial neural networks (ANN), usually simply called neural networks, are computing systems inspired by the biological neural networks. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.

2.LITERATURE REVIEW

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.

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.

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.

3. IMPLEMENTATION

Existing System

Generally there are two type of tumors. One is benign and other in malignant tumor in which benign Tumor is non-Cancer and malignant is a cancer Tumor. Breast Cancer is the most frequent disease as a cancer type for women. There is no system which can give a better performance on the available data.

Limitations:

- ❖ Performance is less
- ❖ Less Accuracy

Proposed Method

In this work, using SVM (Support Vector Machine) and ANN (Artificial Neural Network) to predict breast cancer diseases. First this algorithms will be trained using past disease dataset called 'Wisconsin Breast Cancer', this dataset contains 11 integer values and last value contains class label either 0 or 1, 0 means person is normal and 1 means person is infected with disease. Both algorithms will be trained on previous people's dataset and new person test data will be applied on trained data to predict it class such as 0 or 1.

The effectiveness of ML techniques is compared in term of key performance metrics such as accuracy, precision, recall and ROC area. Both algorithms generate model from train dataset and new data will be applied on train model to predict it class. Based on the performance metrics of the applied ML techniques, SVM has showed the best performance in the accuracy for the diagnosis and prediction from WBC, SVM algorithm is giving better prediction accuracy compare to ANN algorithm.

Advantages:

- ❖ Performance is best
- ❖ Accuracy is more
- ❖ ML techniques can be handled on early detection and prognosis of cancer.

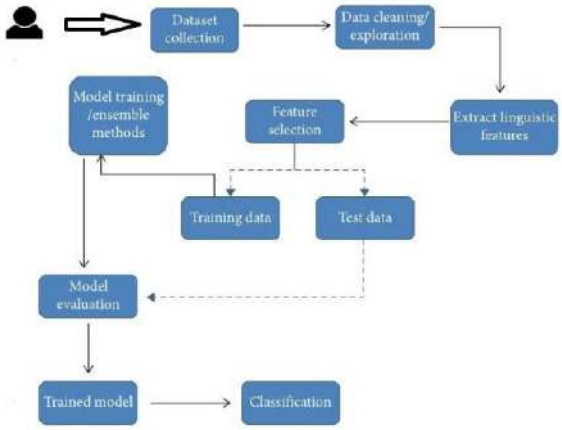

Fig.2: System architecture

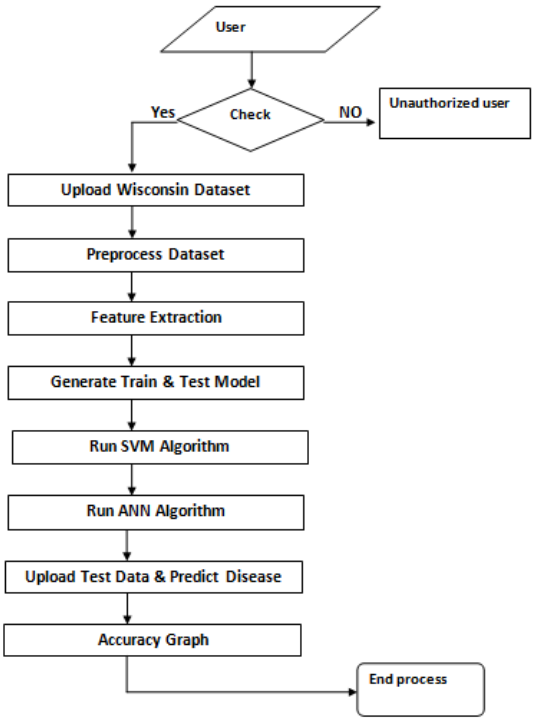

Fig.3: Dataflow diagram

In this work, SVM and ANN two of the most popular machine learning techniques are applied on Wisconsin Breast Cancer (Original) dataset and the result of applied machine learning (ML) techniques are compared according to performance metrics.

4. ALGORITHMS

SUPPORT VECTOR MACHINE (SVM):

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot). The SVM algorithm is implemented in practice using a kernel. The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra, which is out of the scope of this introduction to SVM. A powerful insight is that the linear SVM can be rephrased using the inner product of any two given observations, rather than the observations themselves. The inner product between two vectors is the sum of the multiplication of each pair of input values. For example, the inner product of the vectors [2, 3] and [5, 6] is 2*5 + 3*6 or 28.

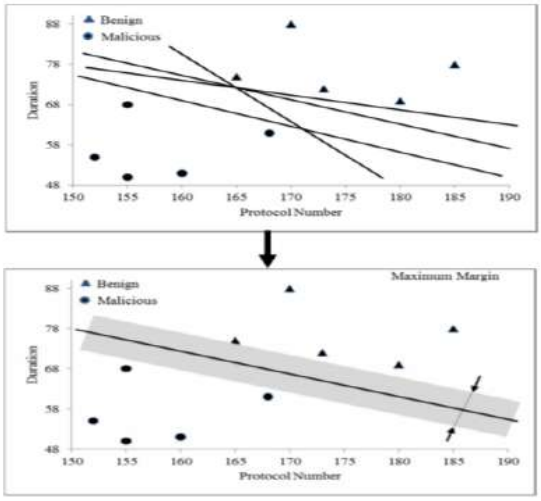

Fig.4: SVM model

ARTIFICIAL NEURAL NETWORK (ANN) :

Artificial Neural Network(ANN) uses the processing of the brain as a basis to develop algorithms that can be used to model complex patterns and prediction problems.

The term "Artificial neural network" refers to a biologically inspired sub-field of artificial intelligence modeled after the brain. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes. Artificial neural network tutorial covers all the aspects related to the artificial neural network. In this tutorial, we will discuss ANNs, Adaptive resonance theory, Kohonen self-organizing map, Building blocks, unsupervised learning, Genetic algorithm, etc.

The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes.

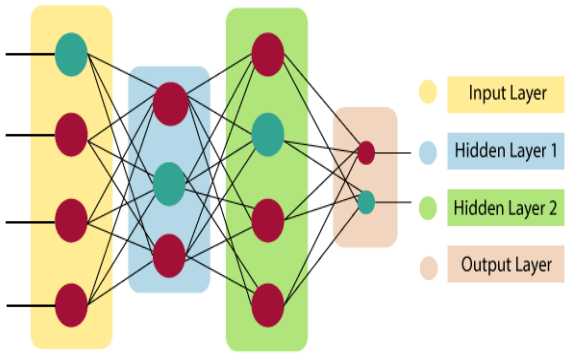

Fig.5: ANN model

An Artificial Neural Network in the field of Artificial intelligence where it attempts to mimic the network of neurons makes up a human brain so that computers will have an option to understand things and make decisions in a human-like manner. The artificial neural network is designed by programming computers to behave simply like interconnected brain cells. There are around 1000 billion neurons in the human brain. Each neuron has an association point somewhere in the range of 1,000 and 100,000. In the human brain, data is stored in such a manner as to be distributed, and we can extract more than one piece of this data when necessary from our memory parallelly. We can say that the human brain is made up of incredibly amazing parallel processors. We can

understand the artificial neural network with an example, consider an example of a digital logic gate that takes an input and gives an output. "OR" gate, which takes two inputs. If one or both the inputs are "On," then we get "On" in output. If both the inputs are "Off," then we get "Off" in output. Here the output depends upon input. Our brain does not perform the same task. The outputs to inputs relationship keep changing because of the neurons in our brain, which are "learning."

5. EXPERIMENTAL RESULTS

Fig.6: Home screen

Now click on 'Upload Wisconsin Dataset' button to upload dataset

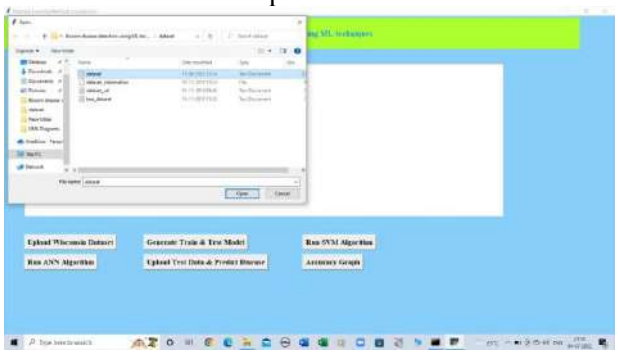

Fig.7 : Upload dataset file

Fig.8: Dataset loaded

In above screen I am uploading 'dataset.txt' file. Now click on 'Generate Train & Test Model' button to split dataset into train and test

Fig.9: Data split into train & test model

In above screen we can see dataset contains total 779 records and training size splits to 545 and test size splits to 234. Now click on 'Run SVM Algorithm' button to generate training model with SVM

Fig.10: Run SVM algorithm

In above screen we can see Accuracy value, precision and recall obtained from SVM. SVM got 66.66% accuracy. Now click on 'Run ANN Algorithm' to get ANN Accuracy

Fig.11: Run ANN algorithm

In above screen ANN got 38% accuracy and SVM got better accuracy than ANN. Now click on 'Upload Test Data & Predict Disease' button to upload test data and to predict disease

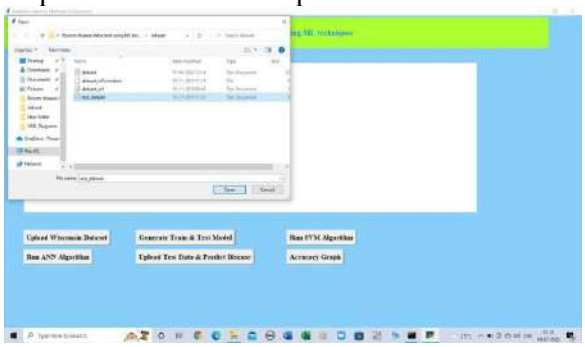

Fig.12: Upload test_dataset file

In above screen I am uploading test_dataset file and below is predicted result

Fig.13: Disease prediction

Now click on 'Accuracy Graph' button to get accuracy graph of both algorithms

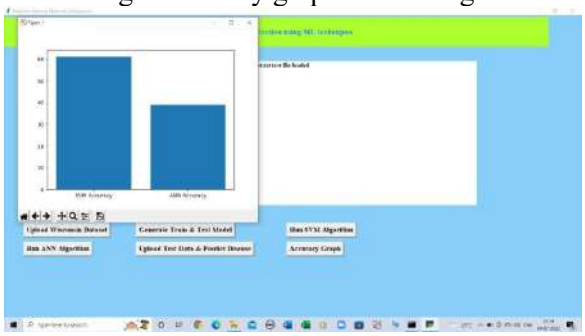

Fig.14: Accuracy graph

In above graph x-axis represents algorithm name and y-axis represents accuracy of that algorithm. From above graph we can conclude that SVM is better than ANN

6. CONCLUSION

Breast Cancer is the most frequent disease as a cancer type for women. Therefore, any development for diagnosis and prediction of cancer disease is capital important for a healthy life. In this paper, we have discussed two popular machine learning techniques for Wisconsin Breast Cancer classification. Artificial Neural Network and Support Vector Machine are used as ML techniques for the classification of WBC (Original) dataset in WEKA tool. The effectiveness of applied ML techniques is compared in term of key performance metrics such as accuracy, precision, recall and ROC area. Based on the performance metrics of the applied ML techniques, SVM has showed the best performance in the accuracy for the diagnosis and prediction from WBC dataset.

7. FUTURE SCOPE

The model can be generalized so that it can predict multiple types of cancer. The accuracy can be raised to 100%, so that it can be more accurate than the doctor's diagnosis. It can improve the domain of oncology, the survival rates can be increased by the early detection and advanced treatments. As we know that, the early cancer can be easily treated with the help of deep learning algorithms like ANN , CNN, RNN and so on we can get more accurate results. Deep learning is subset of machine learning which makes the computation of multi-layer neural network feasible.

REFERENCES

[1] Tolga Ensari, Ebru Aydindig Bayrak and Pinter Kirci "Comparision of Machine Learning Methods for Breast Cancer Diagnosis", 978-1-7281-1013-4, 2021 IEEE.

[2] Ahmed Javid Safi (2019) ,"The Leading Cancer Types in Afghanistan". Journal of Cancer Therapy, 10, 877-881. doi: 10.4236/jct.2019.1011074.

[3] M. Amrane, S. Oukid, I. Gagaoua, T. Ensari, "Breast Cancer Classification Using Machine Learning", Int. Conf. on Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, DOI: 10.1109/EBBT.2018.8391453, April 18-19, 2018.
[4]Siegel, R. L., Miller, K. D., & Jemal, A. (2018). "Cancer statistics, Ca-a Cancer Journal for Clinicians", 68 (1), pp. 7-30.
[5] Min-Wei Huang, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, Chih-Fong Tsai (2017). "SVM and SVM ensembles in breast cancer prediction". PloS one, 12 (1).
[6] Maity, N. G, & Das. S (2017). "Machine learning for improved diagnosis and prognosis in healthcare". In 2017 IEEE Aerospace Conference, pp. 1-9.
[7] Bektas.B, & Babur, S. (2016). "Machine learning based performance development for diagnosis of breast cancer", Medical Technologies National Congress, pp. 1-4.
[8] Konstantina Kouroua, Themis P.Exarchosab, Konstantinos P.Exarchosa, Michalis V.Karamouzisc, Dimitrios I.Fotiadisab (2015). "Machine learning applications in cancer prognosis and prediction", Computational and structural biotechnology Journal, 13, pp. 8-17.
[9] Bazazeh. D, & Shubair. R (2016). "Comparative study of machine learning algorithms for breast cancer detection and diagnosis". In 2016 5th International Conference on Electronic Devices, Systems and Applications, pp. 1-4.
[10] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). "Using machine learning algorithms for breast cancer risk prediction and diagnosis". Procedia Computer Science, 83, pp. 1064-1069.
[11] Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). "Using three machine learning techniques for predicting breast cancer recurrence". J Health Med Inform, 4 (124)..
[12] "UCI Breast Cancer Wisconsin (Original)Datase"t,https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Origina l%29. Last Access: 30.01.2019.
[13] S. Turgut, M. Dagtekin, T. Ensari, Microarray "Breast Cancer Data Classification Using Machine Learning Methods", Int. Conf. on Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, DOI: 10.1109/EBBT.2018.8391468, April 18-19, 2018. 23
[14]Hui-LingChenab, BoYangab, JieLiuab, Da-YouLiuab (2011). "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications", 38 (7), pp. 9014-9022.

Copyright @ 2022 Author