Breast Cancer Prediction System Using KE Sieve Algorithm

G. Priyanka
M.Tech,CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
priyankagummad@gmail.com

V. Rohith
M.Tech,CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
vulli.rohith@gmail.com

Dr.Prasanta Kumar Sahoo
Professor in CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
prasantakumars@sreenidhi.edu.in

Dr.K.Eswaran
Professor in CSE Dept.
Sreenidhi institute of science and technology
Ghatkesar Hyderabad, Telangana-501301
India
kumar.e@gmail.com

***Abstract*:** Breast cancer is a disease which occurs when the cells in the breast grow out of control due to the changes in the genes called mutations. These abnormal cells get accumulated and eventually form a tumor or can be felt as a lump. The main factors causing breast cancer are advancing age and family history. So, earlier detection of breast cancer is necessary to decrease the number of deaths associated with this disease. In this paper, a new non-iterative classifier named KE Sieve is used to detect the presence of cancer by using original Wisconsin Breast Cancer Dataset.

Keywords: *KE's Algorithm, Wisconsin Breast Cancer Dataset.*

1. INTRODUCTION

Breast cancer tumors are of two types, one those are non-cancerous or benign and other one is cancerous which is malignant. If the tumor is benign, cells generally won't spread across other parts of the body tissues and responds well to the treatment. Sometimes benign tumor may become serious if it presses on nearby tissues, blood vessels. In case of malignant tumor, cancer cells grow uncontrollably and spreads across nearby tissues, other parts of the body through lymph nodes[7]. Common factors for cancer are smoking, alcohol consumption. Breast cancer is almost entirely seen in women but men can get too.

Mammogram is an x-ray image of breast, it is used as a screening tool for detecting and diagnosing the breast cancer by doctors [8]. Accurate detection of malignant tumor is a challenging task for many doctors.

Every year, one million women are affected with breast cancer, according to the report of WHO (World health organization) half of them would die, because it's usually late by the time the doctors detect the cancer [3].

Machine Learning (ML) is a field of Artificial Intelligence which is gaining popularity in medical science. It is used for the development of predictive models in order to support effective decision making [1]. It makes the system automatically learn and improve with experience. These ML techniques can reduce diagnostic errors and can better predict whether the cancer is malignant or benign in less time with high accuracy.

The main objective of this paper is to compare most popular Machine Learning (ML) techniques such as RandomForest(RF), Support Vector Machine(SVM) Bayesian Networks (BN),KNN etc. with KE's Sieve algorithm using Breast Cancer dataset. The performance is evaluated in terms of accuracy.

2. LITERATURE SURVEY

There had been many research works done on Wisconsin Breast Cancer dataset, in paper [1], authors **Dana Bazazeh** and **RaedShubair** has stated that the accuracy of SVM on this

dataset is about 97% when compared to Random Forest and Bayesian Networks. Drawback with this approach is SVM doesn't have the property of retraining i.e. if any new point is added they need to train the model again with the previous train points and it also needs several key parameters that need to be set correctly to get good accuracy.

MeriemAmrane and **IkramGagaoua** authors of [3] have briefly explained that KNN gives the accuracy of 97 % with this dataset. Drawback is KNN compares the Euclidean distance with respect to all train points i.e. testing time increases.

Zahra Nematzadeh, Roliana Ibrahim and **Ali Selamat** of [4] have explained that neural networks gives the accuracy of 98 percent on Breast Cancer Dataset. Drawback with this approach is Training or testing takes a lot of time.

3. PROPOSED WORK

3.1 DATASET DESCRIPTION:

Wisconsin Breast Cancer Dataset is taken from UCI machine learning repository [9]. Data was collected by Dr.WilliamH.Wolberg from the University Of Wisconsin Hospitals at Madison, Wisconsin, USA. This dataset consists of 669 instances with 9 attributes, each instance classified as either malignant or benign. Due to the 14 missing values in the dataset, only 683 instances are considered.

Attribute Description:

- Clump Thickness
- Cell Size Uniformity
- Cell Shape Uniformity
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nuclei
- Mitoses

Each of the above attributes are evaluated on a scale of 1 to 10, where 1 is said to be nearest to benign and 10 is nearest to malignant.

3.2. METHODOLOGY:

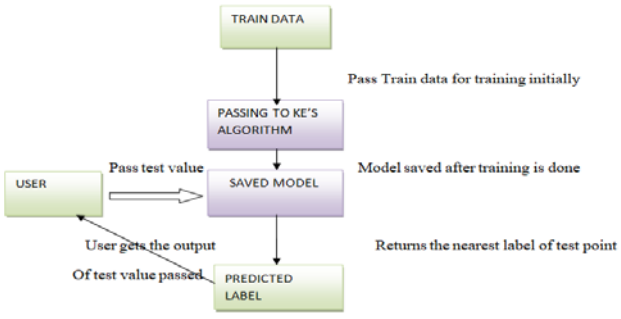

Fig 1: Process Flow Diagram

3.2.1 DATA PREPROCESSING:

We use K-Fold cross validation technique to split the data for train and test samples i.e. if K=10, 9 parts from dataset are for training and 1 part for testing.

3.2.2 KE'SIEVE ALGORITHM:

This algorithm[5,6] is non-iterative and adopts a new approach, which separates N data points of dimension d by hyperplanes. The number of hyper planes needed approximately to separate N data points are $\log_2(N)$ and the computational complexity of this algorithm is approximately $O(d.N.\log_2(N)) + (d^3\log_2(N))$, where N is the data points and d is the dimension of space[2].

The following is the process to perform KE SIEVE algorithm:

Part-1: Training

Step-1: Initially, consider two n-dimensional spaces, in the first space we place the entire N data points and calculate initial planes.

Step-2: To draw the initial planes, for each plane we need to collect the random data point pairs of size dimension (n) and substitute them in this equation $1+\alpha_1x_1+\alpha_2x_2+....+\alpha_nx_n=0$[6]. The planes are drawn in such a way that they pass through the midpoints of data.

Step-3: Now pick each point from first space and move them to second space, which has initial planes and compute the Orientation vectors [6] of all the N-data points.

Step-4: While transferring, if the ov of the current point matches with any other data point's ov then those two points are said to be in same quadrant, they are kept aside in a temporary array and this pair collection is repeated till it reaches the size of dimension (n), these n pairs are used to draw a new hyper plane to separate those points collected.

Step-5: After every hyper plane drawn we need to calculate the ov's again with respect to new hyper plane generated and update the plan coefficients with new hyper plane, this process is continued until each and every train point is covered.

Part-2: Testing

Step-1: Consider the test data and compute the ov's for that corresponding test data with respect to final plane coefficients generated.

Step-2: We calculate dot product for the test ov with all the train ov's and 30 % sampling is done on the dot products.

Step-3: We calculate Euclidean distance of test point with respect to all the sampled trained points and return the nearest label.

Advantages:

- If any new points are to be added for training, we can add them and train the algorithm from where it

has stopped, there is no need of retraining the whole model again.

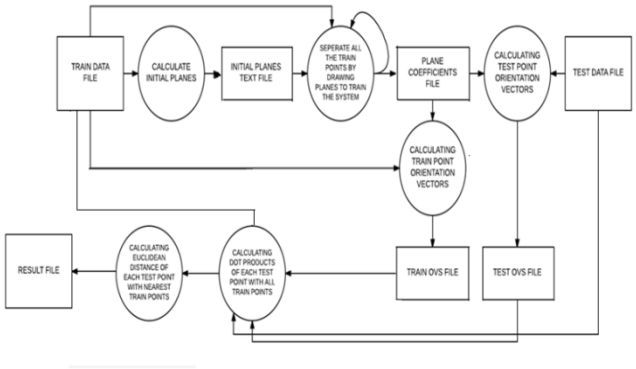

Fig 2:DFT of KE's algorithm

4. RESULTS

This section describes the results obtained when Original Breast Cancer Dataset is given to the KE's algorithm.

- The data for train and test is done using the k-fold cross validation technique, 9 parts for training 1 part for testing.
- 4 Initial number of planes were considered.
- Time taken for training is about 0.29 sec and testing is about 0.01 sec.
- The above result is obtained when, 28 planes are used for classifying data and 30% dot product.
- The accuracy obtained for KE's algorithm is about 98.53 %.
- The confusion matrix obtained for the test data is

Test-Points=68	Predicted (yes)	Predicted (no)
Actual(yes)	55	1
Actual(no)	0	12

5. CONCLUSION

In this paper, "KE SIEVE" is applied on Wisconsin Breast Cancer dataset. The accuracy achieved is 98.53% in the very less time span of 1 second.

ALGORITHM	REFERENCE PAPER	ACCURACY
KE SEIVE	CURRENT PAPER	98.53%
SVM	1	97%
KNN	3	97%

Table-2: Comparison of algorithms in terms of accuracy.

Split ratio of dataset (train-test)	Nearest neighbour [K=1] Accuracy	Nearest neighbour [K=3] Accuracy	Nearest neighbour [K=6] Accuracy
90-10	98.05%	98.45%	98.53%
80-20	95.62%	96.35%	94.89%
70-30	96.09%	97.05%	97.05%
60-40	95.25%	97.44%	96.35%

Table 3: Performance test of KE SIEVE algorithm with different split ratios.

REFERENCE

1. D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016, pp. 1-4.

2. Chitukula, Kavya & Gayatri Likhita, G & Hiranmayi, D & Bantu, Saritha & Eswaran, K. (2017). Classification of Diseased Plants using Separation of Points by Planes. 4. 335.

3. M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4.

4. Nematzadeh, Zahra & Ibrahim, Roliana & Selamat, Ali. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. 1-6. 10.1109/ASCC.2015.7244654.

5. K.Eswaran, "A non iterative method of separation of points by planes in n dimensions and its application" in https://arxiv.org/abs/1509.08742v5 October 23 2015

6. Eswaran, K. (2017). On non-iterative training of a neural classifier Part-I: Separation of points by planes. 10.1109/IntelliSys.2017.8324238.

7. https://www.verywellhealth.com/what-does-malignant-and-benign-meam-514240

8. https://www.healthline.com/health/mammograpy

9. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)