A FRAME WORK TO DETECT BREAST CANCER USING KNN and SVM

RAJESH SATURI[1],
Research Scholar,
K.V. Sai Phani[2]
Assistant professor
Prof.P. PREM CHAND[3],
Professor,

[1,3]Department of CSE,University College of Engineering,Osmania University, Hyderabad-500007, Telangana State, India.
[2]Department of CSE,Santhiram Engineering College, NH-18,Nerawada 'X' Roads,Nandyal, Kurnool, Andhra Pradesh

***Abstract:** The main reason of increasing mortality rate among women is the breast cancer. It makes several hours with the less availability of systems to identify the diagnosis of cancer manually. Hence there is a need to develop an automatic system for early detection of cancer. Several researchers have focused in order to improve performance and achieved to obtain satisfactory results. But unfortunately it will be very difficult to detect the cancer in beginning stages because the symptoms may be inappropriate.Therefore, there is a need to determine and acquire a new knowledge to prevent and minimizing the risk of getting effected with cancer. Machine learning (ML) is algorithms are widely used in detecting breast cancer patterns and predict the grading level. Machine learning techniques can be used to classify the stage of cancer, where machine can be trained from past data and build a model so that it can predict the category of new input.In this paper we used K-nearest neighbors (K-NN) and Support Vector Machine (SVM) on the dataset collected from UCI repository to detect breast cancerwith respect to the results of accuracy the efficiency of algorithm is also measured and compared.*

***Keywords:**Feature selection, Breast cancer, machine learning, KNN,SVM.*

1. INTRODUCTION

Breast cancer, is the second leading source of casualty among women worldwide. Not only among women but according to researcher’s men can also be affected with breast cancer. But majority cause of fatality rate among women is more when compared with men. Many experts from over past few years observing the symptoms of different types of patient’ssymptomsin order to diagnose. But still the accuracy in predicting the cancer is not definite. With the invention of new computing technologies, now its became very easy to acquire huge amount of data and stored.Patientscan perform self inspection by observing the breast and underarm, therefore a patient will be familiar with her breast and can detect the anomalies that observes during her workouts, but it is a time consuming task and involves monthly observations. Diagnosis is the procedure of investigating the stage of cancer as either benign or malignant cases. Supervised machine learning algorithms assigns classes or labels to different object groups and constructs a model to estimate the accuracy of grading the stage of cancer The accuracy of a classifier in general estimated as the percentage of test samples that are correctly classified. Generally, a patient can visit a specialist for mammograms to observe the severity of disease. Image processing techniques can be used to convert that image into any other format, store it in system and extract the features required that gets a more meaningful data where we can apply any machine learning algorithm to detect the stage of breast cancer

There are different types of datasets available,like example Wisconsin Prognostic Breast Cancer Chemotherapy (WPBCC), Wisconsin Diagnostic Breast Cancer (WDBC) and so on. There are different types of machine learning algorithms applied on these datasets that shows different levels of accuracy range between 94.36% and 99.90%.

2. METHODOLOGY

The dataset collected from Wisconsin and implementation is done using python-spyder. In this methodology we applied supervised algorithms and feature selection techniques like Random Forest, chi2, correlation, hypothesis with Dimensionality Reduction technique.

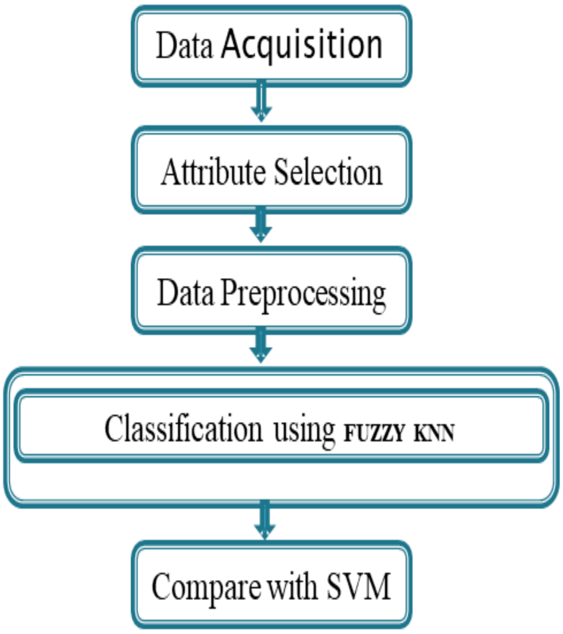


Fig1:Proposed System

3. KNN MODEL:

KNN algorithm stores all the objects and based on similarity measure it classifies new objects into different classes.

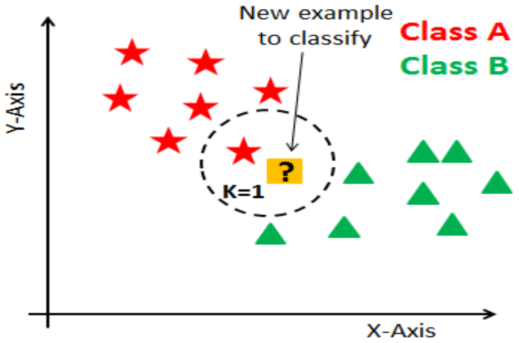


Fig2: KNN

Every objectwill be classified by a majority of votes compared with their neighbors, for example If K = 1, then the object is simply assigned to the nearest neighbor of its class.

Distance functions

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

These are only valid for continuous variables. forinstance, of categorical type of variables, hamming distance can be used. It is a challenging task to standardize the numerical values between 0 and 1 when there is a blend of both categorical and numerical variables in the dataset.

Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

Initially the Optimal value of K can be obtained by inspecting the data. In general, a large K value is more precise as it reduces the overall noise but that's not guarantee. Cross-validation is an another method to determine a better K value by using an independent dataset. Generally, the optimal K value for most of the datasets are between 3 to 10, which produce better results than INN.

4. SUPPORT VECTOR MACHINE CLASSIFIER MODEL:

SVM represents the examples as points in space, so that the examples of the distinct categories are separated by a clear gap as wide as possible. SVMs can perform linear and non-linear classification, SVMs are based on maximum margin linear discriminants, and similar to the probabilistic approaches, but it does not consider the dependencies among attributes.

supervised learning usually involves training and testing data sets which contain data instances. In training dataset each instance contains one target value (class label) and several other attributes (features). The main objective of a classifier is to produce a model that can be able to predict target values of data instances in the testing dataset, without loss of generality, the classification problem can be viewed as a two-class problem in which one's objective is to separate the two classes by a function induced from available examples. The basic approach of SVM classifier is to choose the hyperplane that has the maximum margin.

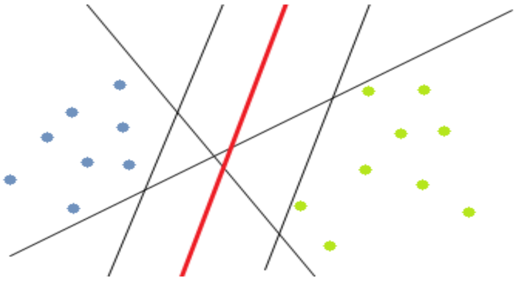
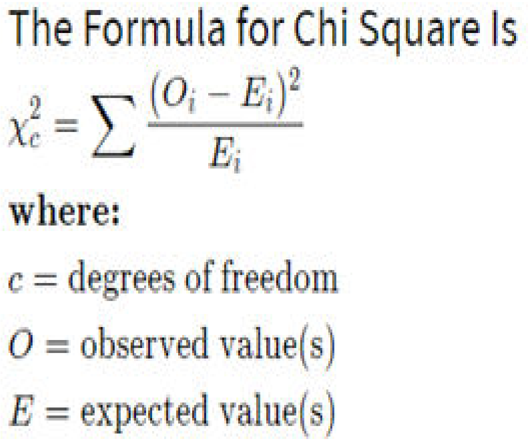


Fig 3: Example of separating hyperplanes

The proposed model is to compare the 2 models to determine the optimized results by making us of the feature selection methods chi2 and correlation & hypothesis.

5. Feature Selection Using Chi^2 Method

In general chi-square test is used in statistics to test the independence of two events. suppose if there are two variables and observed count is O and expected count is E. Chi-Square measures how observed count Oand expected count E deviates each other.



The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom

O = observed value(s)

E = expected value(s)

For example,if we need to determine the relationship between the independent category feature (predictor) and dependent category feature(response). we aim to select the features which are highly dependent on the response.

If two features are independent and the observed count is close to the expected count, then we can consider the smaller chi-Square value. Hence if Chi-Square value is high it indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent.

5.1 Feature selection — Correlation and Hypothesis (P-value)

The linear relationship between two variables, like how close the two variables are related is called Correlation.Generally, the features with highly correlated are more linearly dependent and has almost the same effect on the dependent variable. Hence, when two features are highly correlated, then we can drop any one of the feature.

Probability value (P-Value) or asymptotic significance is a value given for a particular model that, if the null hypothesis is true, a set of statistical observations are greater than or equal to the magnitude observed in results.

In other words, Probability value P-value gives us the possibility of finding an observation under an assumption that a particular hypothesis is true. This probability is used to accept or reject that hypothesis.

6. RESULTS:

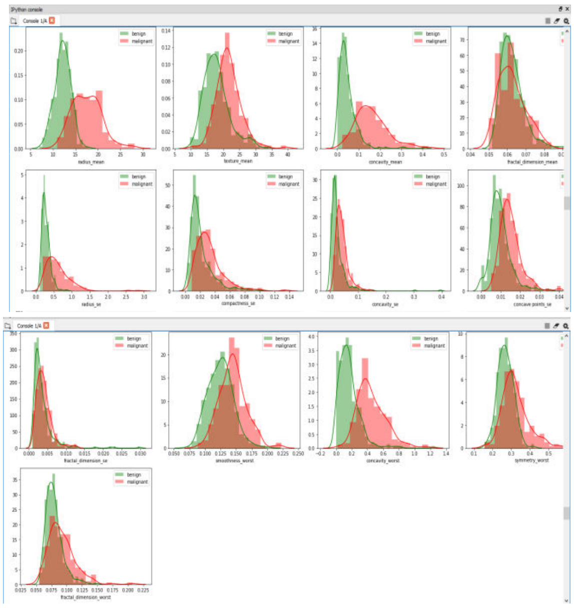


Fig 4 &5 : Data analysis of breast cancer

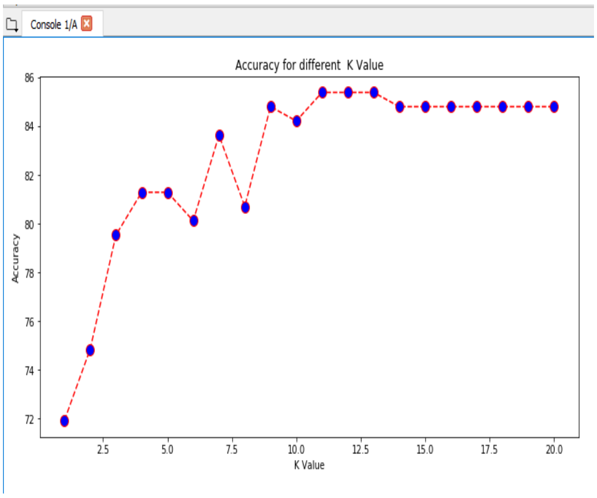


Fig 6Accuracy of KNN

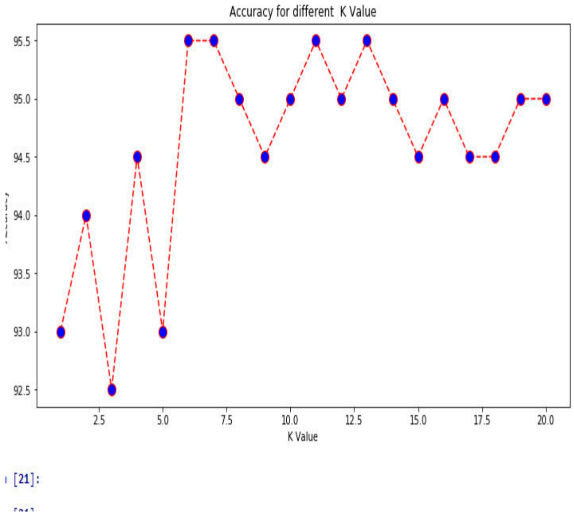


Fig 7: Accuracy of Fuzzy KNN

7. CONCLUSION:

In this paper,the overall methodology, KNN and SVM techniques has given the best results using CHI2 feature selection algorithm over CORRELATION AND HYPOTHESIS algorithm. KNN has given the best results than SVM using CORRELATION AND HYPOTHESIS where as SVM has given best results using CHI2 feature selection.

Result Analysis for train data:

Algorithm	**SUPPORT VECTOR MACHINE**	**K- NEAREST NEIGHBOUR (I=13)**
CHI2	94%	**94%**
CORRELATION & HYPOTHESIS	78%	**84%**

Result Analysis for test data:

Algorithm	**SUPPORT VECTOR MACHINE**	**K- NEAREST NEIGHBOUR (I=13)**
CHI2	96%	**95%**
CORRELATION & HYPOTHESIS	87%	**87%**

REFERENCES

1. Ancy, C.A., Nair, L.S.: An efficient CAD for detection of tumour in mammograms using SVM. In: Proceedings of 2017 IEEE International Conference on Communication and Signal Processing ICCSP 2017, vol. 2018, pp. 1431–1435 (2018).
2. Jothilakshmi, G.R., Raaza, A.: Effective detection of mass abnormalities and its classification using multi-SVM classifier with digital mammogram images. In: International Conference on Computer, Communication and Signal Processing Special Focus IoT, ICCCSP 2017, pp. 1–6 (2017)
3. Johra, F.T., Shuvo, M.M.H.: Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic. In: 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) 2016, vol. 1, pp. 1–5 (2017)
4. Saraswathi, D., Srinivasan, E.: Performance analysis of mammogram CAD system using SVM and KNN classifier. Proc. Int. Conf. Inven. Syst. Control. ICISC **2017**, 1–5 (2017)
5. Ghongade, R.D., Wakde, D.G.: Computer-aided diagnosis system for breast cancer using RF classifier, pp. 1068–1072 (2017).
6. Noor, M.M., Narwal, V.: Machine learning approaches in cancer detection and diagnosis : Mini review machine learning approaches in cancer detection and diagnosis : Mini review (2017)
7. Kanchanamani, M., Perumal, V.: Performance evaluation and comparative analysis of various machine learning techniques for diagnosis of breast cancer. Biomed. Res. **27**(3), 623–631 (2016).
8. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V, Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. **13**, 8–17 (2015).
9. Lg, A., At, E.: Using three machine learning techniques for predicting breast cancer recurrence. J. Heal. Med. Inform. **04**(02), 2–4 (2013).
10. Hussain, M., Wajid, S.K., Elzaart, A., Berbar, M.: A comparison of SVM kernel functions for breast cancer detection. In: Proceedings of 2011 8th International Conference Computer Graphics, Imaging and Visualization CGIV 2011, pp. 145–150 (2011)
11. Rejani, Y.I.A., Selvi, S.T.: Early detection of breast cancer using SVM classifier technique. **1**(3), 127–130 (2009).
12. Selvi, S.T., Malmathanraj, R.: Segmentation and SVM classification of mammograms. In: proceedings of the IEEE International Conference on Industrial Technology, pp. 905–910 (2006).