ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Interactive Dashboard for The Betterment of Patient Health Using Big Data Analytics.

N Gomathi *, Pelluru Srinivasulu, Chakka Venkata Nikhil, and Kakaraparthi Vishnu Vineeth.

Computer Science and Engineering, Vel tech University, Chennai-600062, India.

ABSTRACT

The leading type of cancer that causes a significant worry for many women [1] [2] is breast cancer, which is one of the main cause of death among women globally. The world is experiencing an unpredicted rise of breast cancer cases across all the sections of society. We cannot prevent breast cancer, but we can definitely detect it early and diagnose so that we can achieve longer survival. The reason for the severity of breast cancer arises in the prediction as the doctor cannot predict it accurately. The existing model for prediction is the DT-SVM hybrid model [2] using Hadoop. In our proposed model, we use Random forest-SVM with Pyspark which outperforms the existing model DT-SVM in Accuracy by 7 percent and reduction of Error rate by 2 percent.

Keywords: Breast cancer, data analytics, Pyspark, Random Forest, Accuracy, Error rate, image processing.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Breast cancer arises when the cells in the breast split and grow abnormally. In the last few decades, doctors have made great efforts for early diagnosis to reduce the death rate. This breast cancer can be classified into stages. They are phase 0, phase IA, phase IB, phase IIA, phase IIB, phase IIIA, phase IIIB, phase IIIC and phase IV [3]. Doctors refers phase I to phase IIA as the early stage and from the phase IIB it is the advanced stage. The severity may increase from the early stage to advanced stage either by not detecting the cancer earlier or by not having proper diagnosis. To overcome this detection problem, expert based prediction systems [4] came into existence which is the basic concept of machine learning. These predicting systems must be capable of handling of huge amounts of data from patients suffering from breast cancer. These large datasets for analysis are collected from various local bodies [5] [6] of each and every perspective area. They are: Electronic health care records (EHRS), monitoring or diagnostic instruments, Insurance claims or billings, pharmacies, human resource and supply chain, a real time locating system etc. The existing prediction model is DT-SVM hybrid model using Hadoop. Our proposed model is Random forest-SVM using Pyspark with CHESS software as a front end.

Role of Big Data Analytics in Medical field:

For analyzing and processing the huge amount of medical data, big data analytics [7] comes into light. The life cycle of Big data is the base for all predicting systems like DT-SVM, IBL and SMO. The below diagram gives the clearer vision of the big data life cycle:

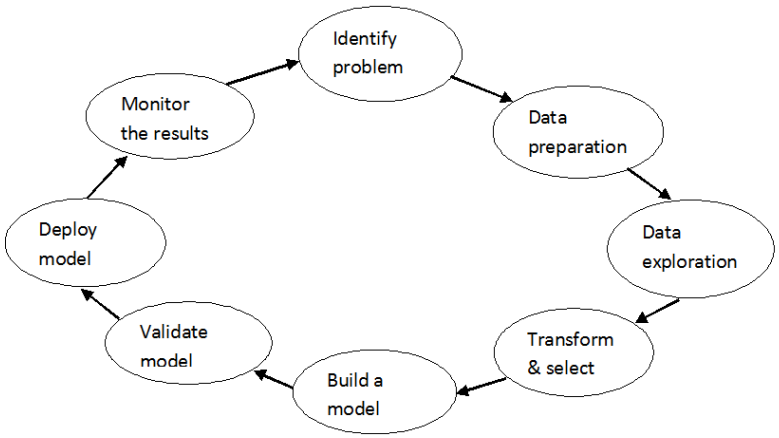

Fig 1: Big Data life cycle

Related works:

In [17] [18] [19], Big data analytics is applied in health care that plays a key role in carrying out authentic time analysis on the sizably voluminous data which gives an opportunity for researchers to utilize advanced techniques and to provide quality health care. It also helps in stimulating a new era of predictive modelling, statistical implements and algorithms to improve clinical tribulation design, analysing disease patterns and more. Although these effects are in the initial stages the future of immensely colossal data analytics is promising for a healthier population with an increase of life expectancy and reduction of health care cost.

In [20] [21] [22] [23], the new advanced techniques for recognition of masses and micro-calcifications of mammograms for further processing are described. There are two ways of enhancing micro calcifications in digitized mammograms such as multifractal approach and modern mathematical morphology. For enhancing masses, methods include image separation and spotting the region of interest which includes both masses and pectoral muscles. There are other methods to detect breast cancer tissues such as Clinical Breast Exam (CBE), Computer-Aided Detection (CAD), Artificial Neural Networks (ANN) etc.

In [24] [25] [26] [27] [28], Image processing using MATLAB stands as a hallmark for depicting the mammogram image in GUI interface to obtain accurate results. The importance of spatial domain processing which consists of gray level transformation and spatial techniques in MATLAB for beginner The use of automotive image processing technique utilizing canny's edge detector is illustrated.

In [29] [30] [31] [32] [33] [34], breast cancer prediction employing data mining relegation techniques such as SVM, Neural networks, KNN, logistic regression, Naïve Bayes and fuzzy logic had been described which includes predictive framework such as WEKA software, hybrid model, algorithms such as Adaboost have been used.

These [35] [36] [37] include current bio-surveillance systems such as BioSense 2.0, health map, Google Flu Trends, EARS, NEDDS, BCON etc., In [38] The different methods and concepts used till date, their functioning with their merits and demerits are discussed.

Mammographic detection of breast cancer:

The main goal of mammography test is the premature recognition of breast cancer. This is done by detection of masses [8] [9] and micro-calcifications [9] which are considered as a consequential designation of breast cancer. Micro-calcifications appear on a mammogram as minute bright tissues of eccentric shape. The possibility of malignant tumours in breast depends upon number, dispersal and nature of micro-calcifications. The cluster composed by micro-calcifications is of two types: homogeneous and heterogeneous. A Homogeneous cluster consists of immense round or oval micro-calcifications that designate a benign tumour. While the heterogeneous cluster consists of diminutive micro-calcifications of anomalous shapes designate high risk of cancer. Masses are another important sign of cancer in the breast. They can occur in various parts of the breast and have dissimilar shapes, dimensions and boundaries. Determination of cancer in the case of the masses depends on the boundaries of their shape. Masses of definite shape, sharp border and if no more masses can be identified within the breast, it indicates the presence of benign, which are not sharp and no definite boundaries cause high risk of cancer (malignant).

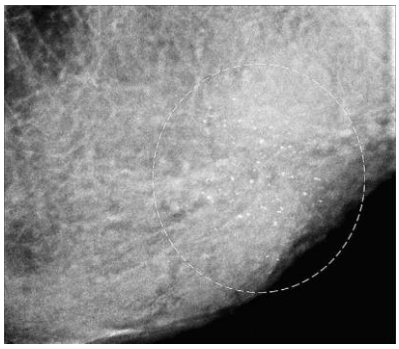

Fig. 2 [10]

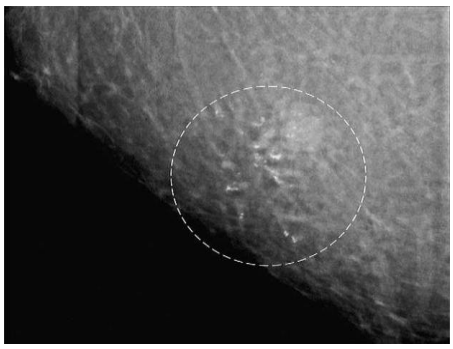

Fig. 3 [10]

Fig 2: shows a mammographic image of the breast region. In this image a collection of micro-calcifications has been noticeable by the white dashed ovals. These are tiny and round which denotes benign cases.

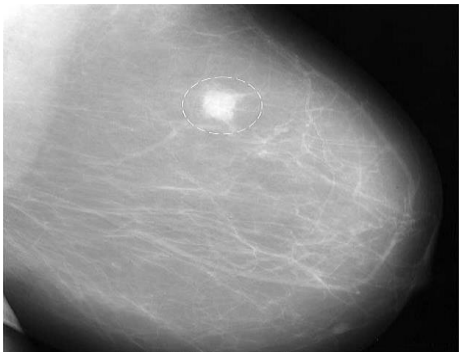

Fig. 5 [10]

Fig. 4 [10]

Fig 3: shows highly heterogeneous micro-calcifications in both shape and size which denotes high risk of breast cancer (malignant ones).

ISSN: 0975-8585

The mass from the fig. 4 has a distinct, sharp border. Moreover, no other masses can be recognized within the breast. These results suggest the lesion is benign. From the fig. 5 the edge of the mass is sharp and is imprecise which denotes that the lesion is malignant.

Apart from finding the disease, radiologists are frequently besieged by large mammograms yielded in the wide-spread screening. Because of this, an important part of visible masses are frequently dropped by radiologists, that cause a great confront in mammographic detection. To tackle this great challenge several computer-aided diagnosis (CAD) [11] approaches have been proposed but these methods had engendered more number of mendacious positive cases. Due to these consequences, Content-based image retrieval (CBIR) [11] techniques came to light, one of the CAD methods which additionally failed when it comes to scalability. So as to extract the abnormalities or features of the mammograms, the image processing technique is utilized. For image processing, haralic [12] algorithm is utilized which can process the image data in multi resolution mode. So the image features can be extracted in an efficient and precise way. The Image processing technique involves the following:

Pre-processing:

In this stage we reduce the speckle noise which condenses the image quality. Speckle noise can be decreased to a greater extent by manipulating the image consequently colour is changed into grey. This is the first and foremost stage in processing a mammogram image.

Filtering process:

This stage aims at reducing the image noise without removing the significant part of the mammogram image like edges, lines, sharp borders of masses which are sensitive in the above areas.

Enhancement process:

It's a process of adjusting the digital image so that more detailed study of the image is possible. Here the underlying features of the mammogram image are extracted that creates the input for further processing.

Feature extraction:

This is the final stage of image processing in which the extracted feature is expected to match with the input data. This feature extraction plays a key role in the image classification using support vector machine algorithm.

Existing model: -

Existing model for breast cancer prediction through the machine learning concept is Decision Tree-Support Vector Machine (DT-SVM) hybrid model using WEKA [2] tool. Classification here has training and testing phases. The main role of SVM is classifying the breast cancer patients either as benign or malignant and that of DT is to optimize the input parameters to the SVM.

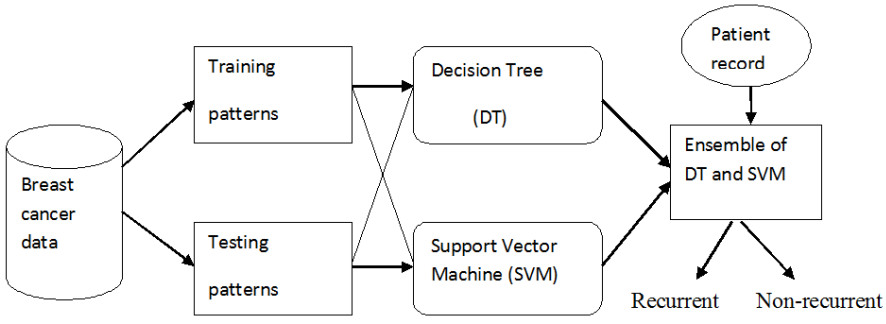

Fig 6: Existing model

ISSN: 0975-8585

Proposed model: -

Our proposed model is Random forest-SVM with Pyspark using CHESS software as the front end and ORBiT as a bio-surveillance system. The existing system for predictive analysis is done by using Hadoop/Map Reduce [13]. We step forward to propose an analytical architecture that consists of "Pyspark" which is faster (100x) than Hadoop/Map Reduce (x) and it also runs faster (10x) on the disk.

Architecture: -

Data collection means all the data sources from the various organizations are given to the system as the input for processing and analysis. Data warehousing is the unit where the data is assembled into a single large warehouse. This large amount of data processing and analyzing is based on the spark streaming model in which data is transferred to the target machine at very high speed. Pyspark is a data processing engine which is used to process and analyze the large amount of data. Data analysis is done by Java, Scala, Python and R. Results are dispersed to lot of servers and reproduced across diverse nodes after examining the breast cancer data. Through these servers a doctor can receive an accurate report of cancer.

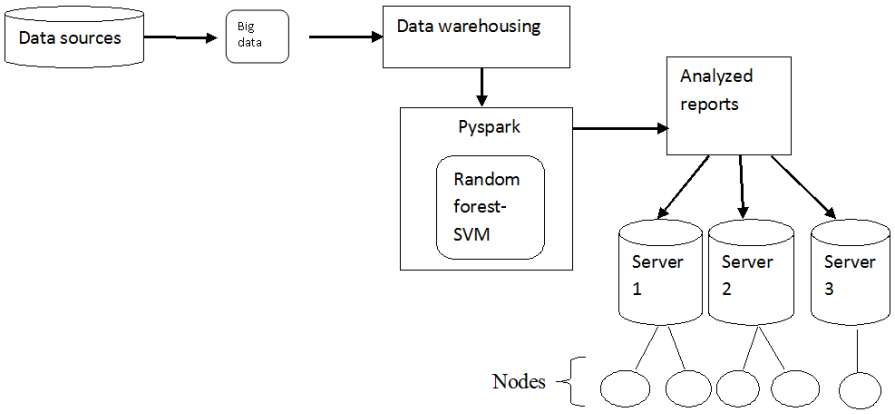

Fig 7: Proposed architecture

To make the data readily available for predictive analysis, first it must be pre-processed. The sample datasets collected from the above data sources [5] [6] may have some erroneous or inconsistent data. This type of data is called as outliers. These outliers may cause incorrect results. So we utilize the concept of data processing for outlier reduction. This outlier reduction can be implemented by using the cluster based outlier detection [14] in which a score is given to the outliers, based on the score they are eliminated from the datasets which is better than traditional distance based outlier detection. After pre-processing the data is sent to training phase, from there the data is fed to the Random forest and SVM module.

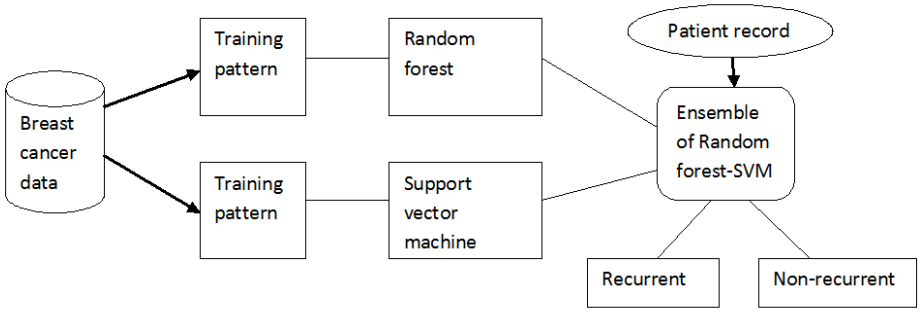

Fig 8: Proposed model

Random forest -Support vector machine: -

ISSN: 0975-8585

This algorithm works on the decision tree basis. It first collects the random sample of data and grows a large number of decision trees which gives more precise and accurate results than a single decision tree. Those are refined and classified to produce the strongest classification model [15]. SVM is the main classifier of breast cancer data which comes under supervised learning which a machine learning task. The key feature of SVM classification is feature selection which identifies the key data sets for analyzing

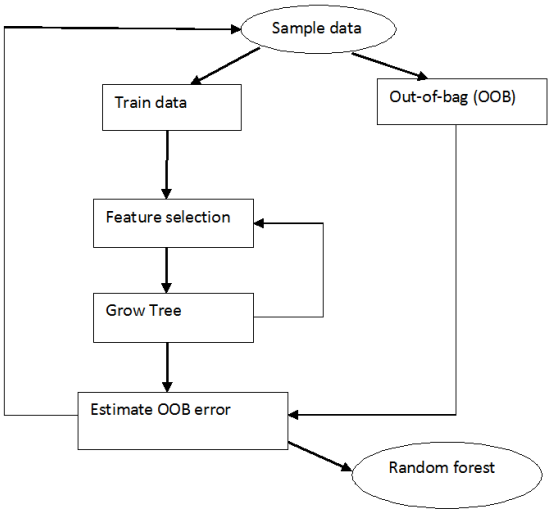

Fig 9: Random forest process

Random forest prototype: -

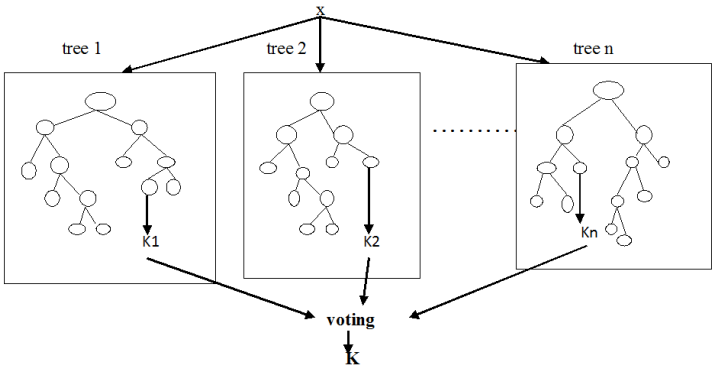

Fig 10: Random forest prototype

Random forest algorithm: -

1. Sketch out 'm' trees from the pounced samples from the breast cancer data.
2. From the pounced samples, do an unsheared gradation tree with these changes at each vertex instead of selecting the best split within all predictions, casually selecting 'm' trees of the predictions and select the best split within these variables.
3. Predict the novel data by augmenting the predictions of 'm' trees (maximum rating for classification, medium rating for regression).
4. A measure of the error rate is got depending upon the training data due to the methods given below:
   (a) For each pounced iteration, predict the root data from the pounced data from the pounced sample (Bremen says Out of Bag data) using the tree which has been built from the pounced sample.
   (b) Accumulate the OOB predictions (based upon the average, every data point is OOB around 36%, so accumulate these predictions) estimate the error rate and say it as the OOB calculation of error rate.

Support Vector Machine (SVM) algorithm: -

Scenario 1: -

(a) Plot every data as a dot in 'm' dimensional space, where m is the set of characters we have with the specification of each characteristic being the specification of particular co-ordinate then the classification is performed by finding the hyper plane that classifies the two classes. SV's are the specified co-ordinates of each and every observation.
(b) Identify the hyper plane, there are three hyper planes a, b and c. Find out the correct hyper plane to classify the two data sets.
(c) Use the thumb rule to pick out the hyper plane.
(d) Choose the hyper plane which separates the two data sets clearly.

Scenario 2: -

(a) The hyper plane a, b and c are separating the classes increasing the distance between closest data point, hyper plane leads to a certain the right hyper plane. This distance is called as margin.
(b) The margin for the hyper plane should be more when compared to other hyper planes because of its robustness.
(c) Choosing a hyper plane with less margin leads to greater choice of misclassification.

Data analysis: - [2]

Based on the domain number of each attribute the patient can be classified into either benign or malignant. The attributes include the clump thickness, the sample code number, uniformity of cell shape, uniformity of cell size, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitosis and class (benign or malignant).

CHESS: -

CHESS is Comprehensive Healthcare Electronic Software System [6] designed to communicate the main server to its large number of respective nodes. It acts as the front end, the following are the things we do with it: we upload the datasets into CHESS, CHESS moves the data to pyspark and users can access their data through a variety of tools such as the tableau which shows the results.

Bio-Surveillance system for breast cancer: -

ORBiT [16] is a data analytics platform for bio-surveillance in the case of breast cancer. It collects the information from all over the country in order to determine the extent of breast cancer. This tool is utilised to generate awareness among the global population by explaining the hazardous circumstances of breast cancer throughout the world. We propose to use this tool in such a way that it acts as a communication bridge between patient and doctor. This tool gives information about a particular patient, such as last visit to the hospital, the stage of cancer, medical prescription and date to visit the hospital again. The patient details are extracted by the pyspark. The tool creates awareness to the patient regarding their health status after their preliminary hospital visit. These data are maintained with the help of central server which integrates each and every hospital and health care centres.

ISSN: 0975-8585

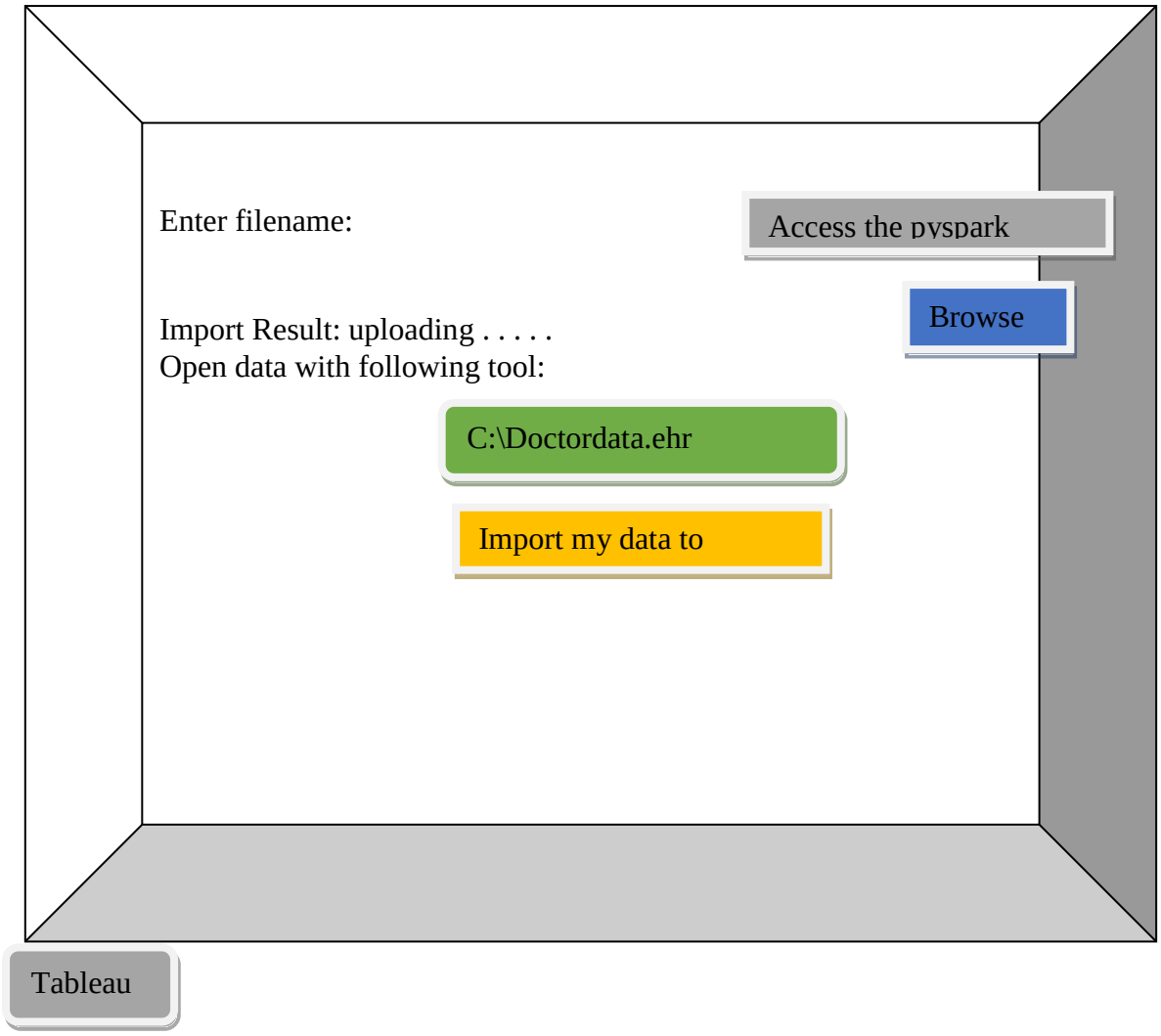

Fig 11: CHESS tool sample interface

ORBiT

Patient name: ABC
Last visit: 21-10-15
Medical prescription: _____
_____
Date to visit again: 06-11-15

Fig 12: ORBiT tool sample interface

CONCLUSION

This paper reduces the breast cancer risk by predicting it early through expert based systems. So that proper diagnosis and proper treatment is given. For this the spark streaming model of pyspark with CHESS

software as the front end in the predictive analysis of breast cancer tissues are used. The cluster based outlier detection and Random forest are used for pattern matching. ORBiT tool kit is used for the breast cancer treatment follow up that has become a deadly disease among women all over the world. The tabular column shows that the Proposed Random Forest with SVM outperforms DT-SVM in Accuracy by 7 percent and reduction of error rate by 2 percent.

Model	Accuracy	Error Rate
Random forest-SVM	98.3%	0.5
DT-SVM	91%	2.58
Naïve Bayes	89%	9.89

Table1: Accuracy and error rate for various classification methods

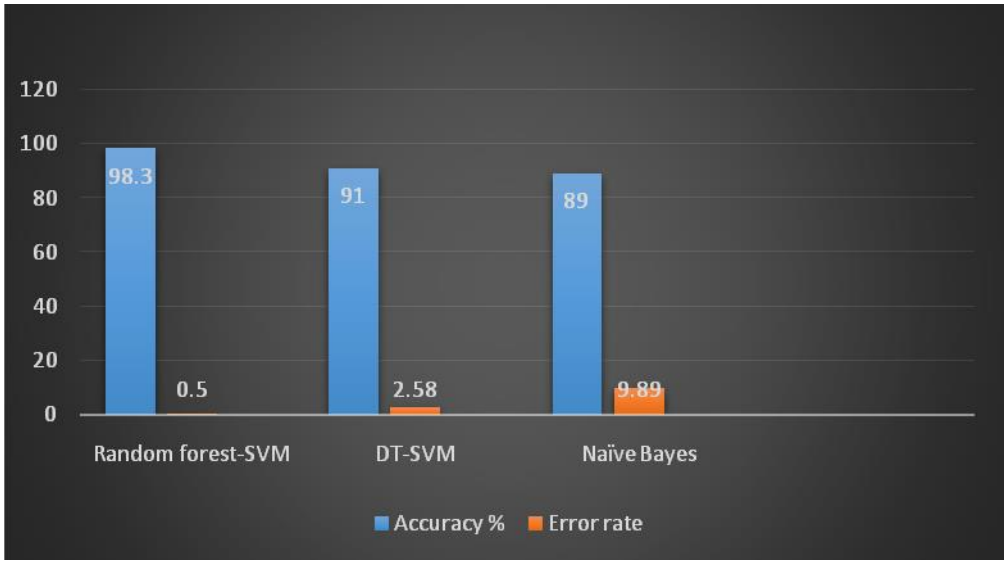

Fig 13: Accuracy and error rate for various classification methods

REFERENCES

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).
[2] K. Sivakami, “Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model”.
[3] American Society of Clinical Oncology. Breast Cancer-Inflammatory: Stages. (http://www.cancer.net/cancer-types/breast-cancer-inflammatory/stages).
[4] Batarseh, F., "Incremental Lifecycle Validation of Knowledge-Based Systems Through CommonKADS", a Doctoral Dissertation Published at the Florida State University Library Services and the Library of Congress, 2011.
[5] Ward, M., Marsolo, K., and Froehle, C., "Applications of Business Analytics in Healthcare", Paper Published at Elsevier's Business Horizons, Volume 57, Issue 5, pp. 571-582, October 2014
[6] F.A. Batarseh, E.A. Latif, Assessing the quality of service using big data analytics with application to healthcare, *Big Data Research* (2015), http://dx.doi.org/10.1016/j.bdr.2015.10.001
[7] A White Paper Published by the SAS Institute Inc., "Using Analytics to Navigate Health Care Reform", 2015
[8] M. LeGal, G. Chavanne, Valeur diagnostique des microcalcifications groupees decouvertes par mammographies, Bull. Cancer (71) (1984) 57–64.
[9] A.C. of Radiology, BI-RADS: Mammography, in: Breast Imaging Reporting and Data System: BI-RADS Atlas. 4th ed., American College of Radiology, Reston, Va, 2003.
[10] Tomasz Arodz, Marcin Kurdziel, Erik O. D. Sevre, David A. Yuen, “Pattern recognition technique for automatic detection of suspicious-looking anomalies in mammograms”.

[11] Menglin Jiang, Shaoting Zhang, Hongsheng Li and Dimitris N. Metaxas, "Computer-Aided Diagnosis of Mammographic Masses Using Scalable Image Retrieval".
[12] Pattern Recognition with measurement and spatial clustering for multiple images (with G. L. Kelly), proceedings of the IEEE, Vol. 57, No. 4, April, 1969, pp. 654-665.
[13] Dr Saravana kumar N M, Eswari T, Sampath P & Lavanya S,"Predictive Methodology for Diabetic Data Analysis in Big Data".
[14] Manish Gupta, Jing Gao, Charu C. Aggarwal and Jiawei Han (2014). "Outlier Detection for Temporal Data: A Survey", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO.9, Pp.2250 – 2267
[15] Derek Kane, Data Scientist – Big Data & Predictive Analytics. Available: www.slideshare.net/DerekKane/data-science-v-decision-tree-random-forests.
[16] Arvind Ramanathan, Laura L. Pullum, Chad A. Steed, Tara L. Parker, Shannon P. Quinn, Chakra S. Chennubhotla, "Oak Ridge Bio-surveillance Toolkit (ORBiT): Integrating Big-Data Analytics with Visual Analysis for Public Health Dynamics."
[17] Manyika, J., Chui, M., Rbown, B., et. al, "Big data: the Next Frontier for Innovation, Competition, and Productivity", Report by the McKinsey & Company, May 2011.
[18] Groves, P., Kayyali, B., Knott, D., and Van Kuiken, S., "The big data revolution in healthcare", Report by the McKinsey & Company's Center for US Health Reform. January 2013.
[19] Archenaa, J., and Anita, M., "A Survey of Big Data Analytics in healthcare and Government" Published at the Proceedings of Elsevier's Procedia Computer Science, Volume 50, pp. 408-413, 2015.
[20] Raman V, Sumari P and Rajeswari M (2010) A theoretical methodology and prototype implementation for detection segmentation classification of digital mammogram tumor by machine learning and problem solving approach. *Int. J. Com. Sci. Issues*. 7 (5), 38-44.
[21] Dubey R, Hanmandlu M and Gupta S (2010) Level set detected masses in digital mammograms. *Indian. J.Sci. Technol.* 3 (1), 9-13.
[22] H. Li, K. Liu, S. Lo, Fractal modelling and segmentation for enhancement of micro-calcifications in digtal mammograms, IEEE Trans. Med. Imag. 16 (1997) 785-798.
[23] L. Cordella, F.Tortorella, M. Vento, Combining experts with different features for classifying clustered micro-calcifications in mammograms, International Conference On Pattern Recognition, vol. 4, 2000.
[24] Acharya T and Ray AK (2005) Image processing: principles and applications. Wiley-Interscience, Hoboken NJ, ISBN 0471719986.
[25] Gonzalez R, Woods R and Eddins S (2004) Digital images processing, 2nd edition, Prentice-Hall, Inc. NJ, 07458, ISBN 0-13-008519-7.
[26] Thaler M and Hochreutener H (2008) Image processing basics using MATLAB.
[27] Vasudevan K, Dharmendra T, Sivaraman R and Karthick S (2010), Automotive image processing technique using cannys edge detector. *IJEST*, 2(7), 2632-2644.
[28] Vij K and Singh Y (2009), Enhancement of images using histogram processing techniques. *Int. J. Comp. Tech. Appl*. 2 (2), 309-313.
[29] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, " An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013.
[30] A.Bellachia and E.Guvan,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.
[31] D. Delen, G. Walker and A. Kadam (2005), Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, vol.34, pp.113-127.
[32] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.
[33] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012,pp. 55-66.
[34] P.Ramachandran, N.Girija and T.Bhuvaneswari, "Health care Service Sector: Classifying and finding Cancer spread pattern in Southern India using data mining techniques", International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 05 May 2012, pp. 682-687.
[35] Efforts to develop a national bio surveillance capability need a national strategy and a designated leader. GAO-10-645, Jun 2010.
[36] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital disease detection -harnessing the web for public health surveillance. New England Journal of Medicine, 360(21):2153–2157, 2009. PMID: 19423867.

[37] J. Lombardo and D. Buckeridge, editors. Disease Surveillance: A Public Health Informatics Approach. John Wiley and Sons, 2006.
[38] Shobhanjaly P. Nair , N. Gomathi , D. Archana Thilagavathy,Big Data Literature Survey, International Journal of Applied Engineering Research. ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 28593-28602