© 2019 IJSRSET | Volume 6 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET196634

A Novel Approach for improving Breast Cancer Prediction Using Wavelet based Feature extraction and SVM

Madhuri Maru[1], Prof. Saket Swarndeep[2]

[1]Information Technology Department, L.J. Institute of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India

[2]Professor, L.J Institute of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India

ABSTRACT

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. Here, a common misconception is that predictive analytics and machine learning are the same thing where in predictive analysis is a statistical learning and machine learning is pattern recognition and explores the notion that algorithms can learn from and make predictions on data. In this paper, we are addressing the problem of predictive analysis by adding machine learning techniques for better prediction of breast cancer. In this, 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 hybrid algorithm in terms of accuracy, precision, sensitivity and specificity.

Keywords : Breast Cancer, Machine learning Algorithms, Image processing, Convolution Neural Network (CNN)

I. INTRODUCTION

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. [17]

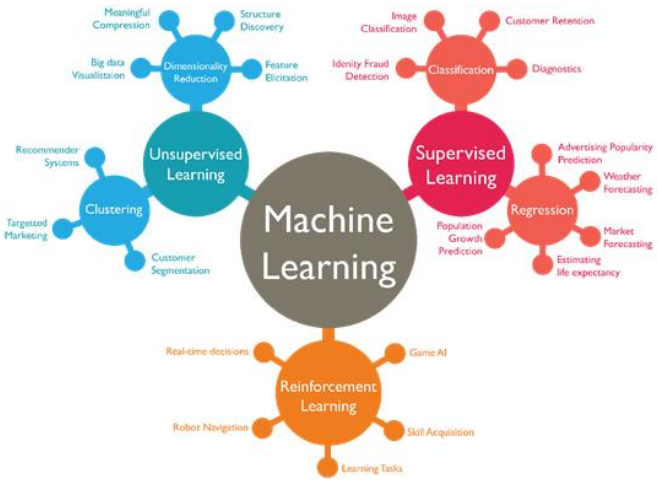

Figure 1. Machine learning methods

The process of learning begins with observations or data, such as examples, direct experience or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide.

Some machine learning methods:

Supervised learning

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans.

You can use supervised learning when the output data is known. The algorithm will predict new data.

Unsupervised learning

In unsupervised learning, an algorithm explores input data without being given an explicit output variable (e.g., explores customer demographic data to identify patterns).

You can use it when you do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you.

Reinforcement learning

It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation.

Application of Machine Learning are in every field, few of them are here listed below.

IMAGE PROCESSING

Image Processing, as its name suggests, is the processing of various algorithms to get the results according to our wish.

There are two types of image processing;

1. Analog Image Processing
2. Digital Image Processing

The analog image processing is applied on analog signals and it processes only two-dimensional signals. The images are manipulated by electrical signals. In analog image processing, analog signals can be periodic or non-periodic. A typical example of analog image processing is a television. In a TV we see pictures due to the controlling of electrical signals ie, using a cathode ray tube. This method of manipulation of images by physical factors like electricity, light etc is called analog image processing.[15]

In digital image processing, digital images are processed digitally. DIP focuses on developing a computer system that is able to perform processing on an image. The input of that system is a digital image and the system process that image using efficient algorithms, and gives an image as an output. Adobe Photoshop and Instagram are the most common examples.[15].

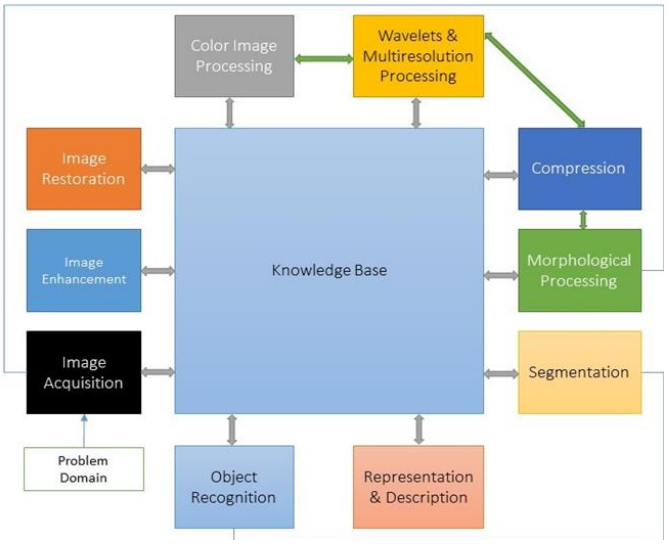

Figure 2. Image processing Block Diagram

II. LITERATURE REVIEW

[1] The objective of the paper was to scale up the machine-learning algorithm that is used for classification by applying each dataset separately and jointly. Mainly 3 types of datasets for the study namely Gene Expression & DM (DNA Methylation) and combination of both for the study they have used. Spark is the big data environment for the fast processing of datasets on different workloads, so that software was used for the analysis work, Weka platform also used to compare the performance with former one. 3 types of classification algorithm namely. SVM, Decision tree and random forest. In the paper work they compared the performance, the efficiency and the effectiveness of the nine predictive models in terms of accuracy, precision, recall, specificity and the area under the ROC on the two platforms to find the best classification.

The result shown that GE data with SVM classification outperforms all other combination, and reaches an accuracy of 99.68%.

In [2] this paper the breast cancer prediction is performed by classifying samples into low or high digital risk score (DRS) groups. The aim of the study was to predict patient outcome based solely on the visual appearance of the breast cancer tissue and be able to reveal complementary and unbiased prognostic information. The classifier was trained using sample images of 868 patients and evaluated and compared with human expert classification in a test set of 431 patients.

Workflow for training and testing the digital risk score (DRS) classification.

- ✓ Feature extraction with CNN
- ✓ Feature pooling with improved finish vector (IFV) and principal component analysis (PCA)
- ✓ Classification with Support Vector machine (SVM)

They concluded that machine learning algorithms together with large scale tumour tissue image series may help approximate the full prognostic potential of tumour morphology.

In [3] this paper they proposed a novel deep learning framework for the the detection of breast cancer in breast cancer in breast cytology using the concept of transfer learning. Deep learning modelled problem specific and gained knowledge during the solution of one problem into another related problem. Proposed work-flow is below,

- ✓ Data pre-processing and augmentation processing
- ✓ Pre-trained CNN Architecture for feature extraction using GoogleNet, VGGNet, ResNet
- ✓ Transfer learning.

GoogleNet, VGGNet and ResNet architecture individually gives average classification accuracy of 93.5%, 94.15%, 94.35% respectively, while the proposed framework gives an accuracy of 97.525% so they concluded that this framework gives excellent results regarding accuracy without training from scratch which improves classification efficiency.

In [4] this paper is to categorise and automatically classify stromal regions according to their maturity and shown that the classification agrees with that of skilled observers, hence providing repeatable and quantitative measure for prognostic studies. They use single and multi scale basic image feature and local binary binary patterns, In combination with a random decision trees classifier for classification of breast cancer storma regions-of-interest (ROI).

The work demonstrated the ability of texture-based image analysis to differentiate breast cancer storma maturity in clinically acquired H&E stained slides at least as well as skilled observers. The performance of single & multi scale RF classifiers were evaluated on the test set of 52, out of these 15 represented mature storma and 37 immature storma.

In [5] this paper they proposed the new algorithm for the weighted KM-SVM to improve accuracy, to accommodate higher error rate of KM-SVM they suggested weighted KM-SVM and evaluated performance of each of classifiers through various experimental scenarios. Notice that the weighted KM-SVM (= 0.278) (and SVM = 0.265) considerably improves the non-weighted KM-SVM (= 0.291) (and SVM = 0.26) They concluded that the proposed weighted KM-SVM is effective to diminish its error rates.

COMPARIOSION TABLE

PAPER	PUBLICATION & YEAR	METHOD	RESULT ACCURACY	FUTURE WORK
1	IEEE-2019	ML + BIG DATA (Spark platform)	99.68 %	Improve the performance of these classification using balanced dataset and feature selection.
2	SPRINGER – 2019	ML + CNN +Tissue Microarray (TMA)	0.60 (Accuracy of DRS grouping)	Future studies required for validate findings test similar algorithm on larger data sets.
3	ELSEVIER - 2019	Deep learning framework +CNN architecture+ transfer learning	97.52 %	In future, both hand-crafted features along with CNN features to improve classification accuracy.
4	IEEE - 2017	Multiscale image feature+ LBP+Random tree	0.86 (accuracy of lightlines filter with LBP)	The investigation of the optimal image size for this particular problem, this would also indicate the scale at which changes
		classify.		in maturity could be detected.
5	KimSpringer plus-2016	Weighted KM-SVM	-	New weighting scheme in proportion to size of clusters to improve more in accuracy.

III. METHODS

1. Support-vector machine

In machine learning, support-vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

2. Random Forest

Random forests or random decision forests are an ensemble learning methods for classification , regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

3. Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

4. Convolutional neural network

In deep learning, a convolutional neural network is a class of deep neural networks, most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks, based on their shared-weights architecture and translation invariance characteristics.

IV. CONCLUSION

Survey of research papers give me an insight of techniques and algorithms used in the prediction of breast cancer. A performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Random Forest on the Breast Cancer datasets are conducted. Main parameters for the comparison were Accuracy, precision, Sensitivity, Specificity. By using mentioned techniques in the proposed system will definitely help in better prediction of breast cancer.

V. REFERENCES

[1]. SARA ALGHUNAIM, On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context (IEEE 2019).

[2]. Riku Turkki, Breast cancer outcome prediction with tumour tissue images and machine learning (August 2019, Volume 177, Issue 1,)

[3]. SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din,Joel J. P. C Rodrigues, A Novel Deep Learning based Framework for the Detection and ClassiPcation of Breast Cancer Using Transfer Learning, Pattern Recognition Letters (2019), doi:https://doi.org/10.1016/j.patrec.2019.03.022M. Veta et al., “Breast cancer histopathology image analysis: A review,” IEEE Trans. Biomed. Eng., vol. 61, no. 5, pp. 1400–1411, May 2014.

[4]. Automated Classification of Breast Cancer Stroma Maturity From Histological Images (IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 64, NO. 10, OCTOBER 2017).

[5]. Weighted K-means support vector machine for cancer prediction(Kim SpringerPlus (2016) 5:1162, DOI 10.1186/s40064-016-2677-4).

[6]. A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in Data Mining Techniques for the Life Sciences (Methods in Molecular Biology), vol. 609. Clifton, NJ, USA: Humana Press, 2010, pp. 223_239. doi: 10.1007/978-1-60327-241-4_13.

[7]. K. P. Murphy, Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2012.

[8]. International Agency for Research on Cancer (Iarc) And World Health Organization (Who). Globocan 2018: Age Standardized (World) Incidence and Mortality Rates, Breast. Accessed: Sep. 1, 2018. Online].Available:https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf

[9]. A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in Data Mining Techniques for the Life Sciences (Methods in Molecular Biology), vol. 609. Clifton, NJ, USA: Humana Press, 2010, pp. 223_239. doi: 10.1007/978-1-60327-241-4_13.

[10]. Wang J, Wu X (2005) Support vector machines based on K-means clustering for real-time business intelligence systems. Int J Bus Intell Data Min 1, 1.

[11]. B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Syst. Appl., vol. 41, pp. 1476_1482, Mar. 2014.

[12]. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition. pp 770–778.

[13]. M. M. Fraz et al., Retinal Vessel Extraction Using First-Order Derivative of Gaussian and Morphological Processing. Berlin, Germany: Springer, 2011, pp. 410–420.

[14]. J.-M. Chen et al., "New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images," Sci. Rep., vol. 5, pp. 1–13, May 2015.

[15]. Circuits Today (http://www.circuitstoday.com/what-is-image-processing)

[16]. Automatic MRI Breast tumor Detection using Discrete Wavelet Transform and Support Vector Machines (https://ieeexplore.ieee.org/abstract/document/8909345/)

[17]. GeeksforGeeks (https://www.geeksforgeeks.org/introduction-machine-learning/)

Cite this article as :

Madhuri Maru, Prof. Saket Swarndeep, "A Novel Approach for improving Breast Cancer Prediction Using Wavelet based Feature extraction and SVM", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 6 Issue 6, pp. 113-118, November-December 2019. Available at doi : https://doi.org/10.32628/IJSRSET196634 Journal URL : http://ijsrset.com/IJSRSET196634