Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 593–601

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Breast Cancer Prediction using varying Parameters of Machine Learning Models

Puja Gupta[a], Shruti Garg[a]*

[a]*Birla Institute of Technology, Mesra, Ranchi, 835215, India*

Abstract

Malignancy of tumor has caused major number of deaths among women. Machine learning tools with proper hyper parametric can help in identifying tumors efficiently. This paper presents six supervised machine learning algorithms such as k-Nearest Neighborhood, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine with radial basis function kernel. Deep learning using Adam Gradient Descent Learning was also applied because it combines the benefits of adaptive gradient algorithm and root mean square propagation. A unique hyper parametric change in each model is shown so that it gives better accuracy within the model as well as comparing each model with one other. The result of deep learning as the most accurate with minimum loss. The accuracy achieved by deep learning using Adam Gradient Descent Learning is 98.24%.

Keywords: WBDC; Adam, ML, Deep Learning

1. Introduction

Breast cancer is diagnosed when a malignant tumour is found in the breast tissue. A malignant tumour is a type of tumour that spreads into nearby cells or even around the whole body. Breast cancer can affect both men and women, but is more common in women [1]. In 2018, over two million new breast cancer modalities were estimated by World Cancer Research Fund, of which 626,679 modalities were fatal cases. Among different types of oncology cases, 11.6% were breast cancer cases with 24.2% of those cancers affecting women [2].

A sign of breast cancer is any new hard mass or lump in breast tissue. However, not all lumps are cancerous. Cancerous lumps can be identified through mammography. Only 78% of women are diagnosed with cancer correctly through mammograms [25]. Thus, there are several other methods in use for the diagnosis of breast cancer.

* Corresponding author. Tel.: +91-9430730791.
E-mail address: gshruti@bitmesra.ac.in

10.1016/j.procs.2020.04.064

Digitised fine-needle aspiration (FNA) of a breast mass is a method that is used in the Wisconsin breast cancer dataset (WDBC) [16]. Through this test, many features such as radius, texture, area, etc. were measured from the cells extracted from lumps. There are two classifications of cells that can be found: malignant (cancerous) and benign (non-cancerous). The malignant cells grow into surrounding tissues and it can also spread to other areas of the body. Benign cells do not attack nearby tissues and do not spread to other parts of the body. However, benign cells can be dangerous if they connect to vital structures in body such as blood vessels or nerves [3]. The accuracy of cancer detection by experienced physicians is 78%, as previously stated, while machine learning techniques can provide accuracy of up to 97% [4]. Thus, supervised machine learning (ML) techniques are applied here for the diagnosis of malignant or benign cells. Although machine learning was also applied by many researchers in the past, a parametric study of ML methods is still lagging. Thus, hyperparameter selection in various machine learning models has been done here.
Table 1 shows related studies done in past on WBCD dataset:

Table 1. Literature Review

Model	Accuracy	Reference no. & Year
GRU-SVM, Linear Regression, Multilayer Perceptron, KNN, SoftMax Regression.	99.04%	[5], 2018
SVM, Naïve Bayesian, KNN	97.13%	[6],2016
LDA infused with SVM	98.82%	[7],2019
SVM and relevance vector Machines	99.28%	[8],2018
Bayesian, Decision Trees , ANN & SVM	97.23%	[9],2015
Decision trees	70%	[10],2017
Association Rules AND Neural Network.	95.6%	[11],2004.
Naïve Bayesian	90.41	[12],2016
Ensemble Method	89.2%	[13],2017
Relevance Vector Machines, SVM, Neural Network	98.4%	[14],2013
Radial Bias Function Neural Network (RBFNN)	99.59%	[15],2018

M. Angrap used six machine learning algorithms to classify tumours cells. A variant of long short term memory neural network was implemented by them which is called Gated Recurrent Unit (GRU). The softmax layer of neural network was replaced by linear support vector machine (SVM) [5]. The accuracy of GRU SVM was highest shown in table 1. H. Asri et al. were used Waikato Environment for Knowledge Analysis (WEKA) tool to apply data mining algorithms, providing the best results with SVM [6]. Omondiagbe et al. designed a computer aided design system through fusion of linear discriminant analysis with SVM in a reduced dataset to give 98.82% accuracy [7]. M Kumari et al. used a unique method via machine learning tools to find unknown patterns in datasets using ML algorithms to justify the best prediction [8]. K. Kourou et al. used a different type of cancer dataset, specifically

WBCD, for breast cancer prediction using four machine learning tools [9]. Karabatak et al. applied association rules along with neural network in order to train the model and after that cross validation was applied to increase accuracy of 95.6% [11]. A naïve bayes classifier with a new weight change approach was applied in [12]. Mohebian et al. studied the prediction of the recurrence of cancer using ensemble learning [13]. Gayathri et al. surveyed three ML models with best results using a relevance vector mechanism [14]. Payam et al. applied data pre-processing along data reduction with a radial basis function network (RBFN) classification technique to get optimal results [15].

A classifier using different machine learning models can predict equally well, as shown in Table 1. Thus, the selection of an appropriate model is difficult, which is why a parametric study of different models has been performed here. Along with machine learning, an optimal technique of deep learning has also applied for classification. Deep learning was applied using the Adam gradient descent learning technique through an artificial neural network. The uniqueness lies in the implementation as well as parameter utilisation in each model.

The rest of paper organized in different sections. Section 2 describe about dataset, pre-processing of data is explained in section 3. Section 4 briefly describes machine learning algorithms implemented in this work. Section 5 describes results and discussions of experiments and conclusions is set out in section 6.

2. Data-set

The data used in present studies to carry out the experiments are taken from Wisconsin Breast Cancer Dataset (WBCD), which is already labelled as malignant and benign. The dataset consists of 30 features computed using fine-needle aspiration (FNA) of the breast mass. Cancer datasets are normally in the form of images. The database consists of feature vectors describing the cell nuclei of image. The attributes of cell nuclei consist of ten real-valued features. These features are described as:

a) radius: mean of distances from centre to points on the perimeter, b) texture: standard deviation of grey-scale values, c) perimeter, d) area, e) smoothness: local variation in radius lengths, f) compactness= perimeter^2 /(area - 1.0), g) concavity: severity of concave portions of the contour ,h) concave points: number of concave portions of the contour, i) symmetry, j) fractal dimension= (“coastline approximation" - 1).

The values of features are in four significant digits. There are total 569 records available in which 357 are benign and others 212 are malignant [16].

3. Dataset Pre-processing

In order to counter irrelevant assignment, the dataset was standardised using the equation (1).

$$Z = \frac{X-\mu}{\sigma} \tag{1}$$

where, X represent features to be standardized, μ is mean and σ is standard deviation. The StandardScaler().fit_transform() of python was used to do standardization of data[17].

4. Machine Learning (ML) Algorithms

The machine learning algorithms were used in this paper for the classification of malignant and benign tumour cells. This paper includes the parametric study of six different machine learning algorithms. The short description of ML methods used in this paper is given in subsequent subsections along with their principal parameters.

4.1 K-Nearest Neighbour (KNN)

K-nearest neighbour is the non-parametric lazy algorithm. The nearest neighbours are selected based on Euclidean distance calculated between x and y vectors given in the equation (2). The result of KNN varies for different values of K [21]. A large value of K will cause overlapping in classes, while a smaller value of K increases computations.

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

4.2 Logistic Regression (L_G)

Logistic regression was found to be suitable for this problem because there are only two classifications in this work. Logistic regression can be applied using two parameters – L1 and L2 – and only the L2 parameter is applied here because it considers all feature vectors as important. The L2 regression is also called ridge regression and can be calculated by the given regularisation formula in the equation (3), which is an estimation of sum of square errors, and can also specify the constraints.

$L2(C)=w^*=\text{argmin}\sum_i 1n[\log(1+\exp(-z_i))] + \lambda^*\sum(w_j)^2$ (3)

Where, $\sum (w_j)^2$ is a regularization term,

$\sum [\log (1+\exp (-z_i))]$ is the Loss term.

λ is a hyper parameter.

'C'=coefficient of regularization is used a$\sum (w_j)^2$ is a regularization s a parameter [18].

4.3 Decision Tree (DT) Type equation here.

A decision tree uses a tree-like model of decisions and their possible outcomes. The main algorithm of a DT is called Iterative Dichotomiser (ID3), which uses Entropy *or* Information Gain of each attribute to construct the decision tree [22]. The parameters of a decision tree used here for tuning are max-depth, min-samples-leaf, and max-leaf-nodes.

4.4 Random Forests (RF)

Random forests are an ensemble method for categorisation, regression and differentiation of work that are operated by construction of decision trees during training time. Random forests are used to resolve problems of overfitting in decision trees [19]. The random forests are collections of trees, and the final decision is taken based on a majority vote, as shown in Figure 1.

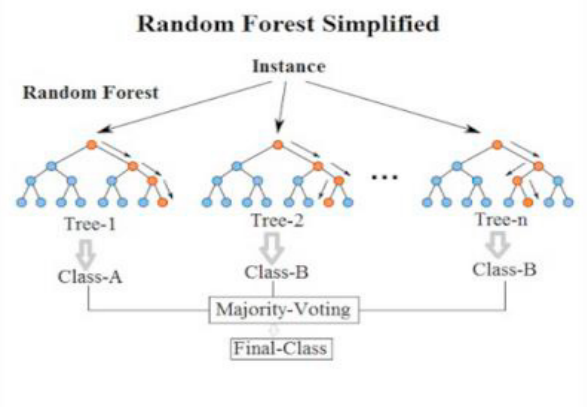

Fig. 1. Random Forests showing different decision trees.

Parameters used in random forests are n_estimators and define the number of trees to be used in the forest. Since a random forest is an ensemble method for the creation of multiple decision trees, the hyperparameter used to control the number of trees are:

1. max_features (number of attributes to be selected from data for randomisation)

2. max_depth (for pre-pruning of trees)

3. max_features=sqrt(n_features) (for classification).

4.5 Support Vector Machine (SVM)

An SVM is an ML algorithm based on constrained minimisation problems. To find the maximum separation distance between objects the dot products of support vectors and the objects is to be calculated. The idea is to map the largest margin between the classes [23]. The concept lies in the conversion of the non-linear separable dataset into a better dimensional space where a hyperplane can be found that separates the objects. The kernel trick used in present work is a radial basis kernel shown in equation (4).

$$K(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2) \quad (4)$$

where,

$K(x_1, x_2)$ is radial bias equation for points or region (x_1, x_2) and γ is spread of kernel. A low value of γ leads to low decision boundary whereas high values of this parameter gives higher decision boundaries.

4.6 Deep Learning Using Artificial Neural Networks

Deep learning is a higher version of implementation of machine learning algorithms. Deep learning introduces non-linear properties to artificial neural networks. Because the present problem is supervised classification, a backpropagation algorithm has been used to train neural networks along with Adam Gradient Descent cost function. The Adam learning combines adaptive gradient algorithm (AdaGrad) and root mean square propagation (RMSProp). Details of Adam is found at [26]. Parameters used in deep learning are as follows:

1. Activation function is ReLU. described in equation (5)

$$f(x) = \max(0, x) \quad (5)$$

2. Sigmoid function is applied after ReLU to convert the output into two classes

3. Batch size depends on the number of objects that are going to be propagated through the network.

4. Epoch is a complete pass through all the training data. The system performance is evaluated using binary cross-entropy loss and increases the predicted probability diverts from the actual value [20].

5. Results and Discussion

The experiment has been performed in 32GB RAM on an intel core i5 8th generation processor with Jupyter notebook of python 3.0. The scikit-learn library functions such as pandas, matplotlib, TensorFlow, and Keras have been used for the experiment.

The evaluation of different models has been done by statistical measures such as accuracy, precision, recall, and F1 score calculated by the equations (6)-(9):

Accuracy = (True Positive +True Negative)/(TP+FP+FN+TN) (6)

Precision = True Positive/(TP+FP) (7)

Recall = True Positive/(TP+FN) (8)

F1 Score = 2*(Recall * Precision) / (Recall + Precision) (9)

where, TP stands for true positive, TN stands for true negative, FP is called false positive rate and FN is called false negative rate taken from confusion matrix.

5.1 K- Nearest Neighbour

Experiments has been done for 3 different values of K shown Table 2. Best accuracy is found at K=6.

Table 2. Accuracy while changing parameter 'K'

Neighbours	K=3	K=5	K=6
Accuracy	.890	.930	.958

Figure 2 shows class distribution of benign and malign class on test and train subset and the outcome is predicted using majority voting system [24]. In majority voting system the neighbours will vote for particular class.

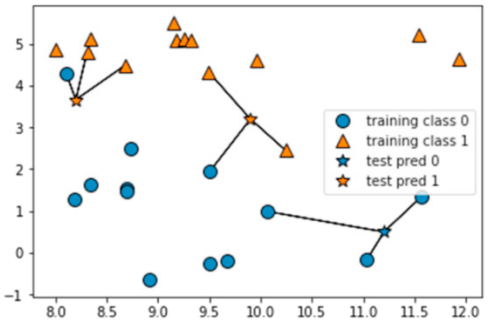

Fig.2. Class distribution by KNN model.

5.2 Logistic Regression(L_G)

Logistic Regression has been applied for three different regularization parameter 'C' shown in table 3. Best accuracy found at C=100.

Table 3. Result evaluation based on 'C' parameter variation

Coefficient Value(C)	C=.10	C=1	C=100
Accuracy	.930	.958	.965

5.3 Decision Tree (DT)

The accuracy of decision tree has been predicted for 2 different depths shown in table 4. Highest accuracy achieved at depth 4.

Table 4. Result of decision tree for different depth

Depth	Normal_Depth=2	max_depth=4
Accuracy	.953	.958

5.4 Random Forest (RF)

The parameters to take in account is n_estimators(number of decision trees).

Table 5. Result of random forest for different n_estimator

n_estimators	10	100	150
Accuracy	.890	.972	.950

N_estimator 100 gives best accuracy shown in table 5. Hence number of decision tree selected as 100.

5.5 Support Vector Machine (SVM)

'C' as regularization parameter with radial basis functions.

Table 6. Result of different parametric results

Parameter	C=0	C=1.0	C=100
Accuracy	.629	.951	.972

The 'C' value is evaluated from C=0, C=1.0 and C=100, the best result found at C=100 as 97.2% shown in Table 6.

5.6 Deep Learning Using Artificial Neural Networks Learning

The result of deep learning varies for different number of epochs. Neural network is trained here for epoch =1 to 150 to increase accuracy and decrease loss as shown in Table 7.

Table 7. Result of deep learning for different epochs

DL_ANN	Epoch=1	Epoch=150
Accuracy	.5605	.9902
Loss	.6928	.0419

After selecting the best hyper parameters for different models. The classification is done for best selected hyper parameter and accuracy, precision, recall and F1-score has been calculated which is shown in Table 8

Table 8. Statistical metrics evaluation score for different machine learning models

Model	Accuracy	Precision	Recall	F1-score
KNN	95.8%	93.5%	93.5%	93%
L_G	95.8%	96.5%	95.0%	95.5%
DT	95.8%	96.5%	95%	95.5%
RF	97.2%	97%	97.5%	97%
SVM	97.2%	97.5%	97%	97%
DL_ANN	98.24%	98%	98%	98%

Six different ML models were applied, and their results are discussed above. Each model gives a vivid result of accuracy, recall and f1-score, which differs between all models. The highest level of accuracy achieved was by deep learning with ANN, with a score of 98.9%. The Adam gradient descent learning tries to minimise errors as well as train data to maximum efficiency and this provided the best result. The second highest result achieved by SVM and random forest was 97%. The accuracy of KNN and logistic regression was the lowest, which is deemed unacceptable. Although SVM and random forest have given equal accuracy, both have their own characteristics. Random forest doesn't present the problem of overfitting, but was found worst for high dimensional sparse data, whereas SVM is adaptable for complex and higher dimension datasets. They can also be applied for linear and non-linear data. Considering the above advantages of SVM, this method proved to be a more superior model.

6. Conclusion

The paper presents the working of six machine learning models by utilising their hyperparameters for the Wisconsin Breast Cancer Dataset. A supervised classification of malignant and benign cells has been done by different machine learning algorithms as well as by deep learning. The accuracy found by Adam Gradient Learning is highest because it combines benefits of AdaGrad and RMSProp. AdaGrad is suits to computer vision problems and RMSProp works well for nonstationary signals. The rectified linear unit (ReLU) function is used here, which did not cause a vanishing gradient problem and allowed the model to learn faster and perform better. This work can further be extended for breast cancer classification using medical images as these play an important role in the diagnosis of cancer.

References:

[1] India against cancer 2019, "*Breast Cancer*", National Institute of Cancer Prevention and Research, viewed 12 November 2019, <http://cancerindia.org.in/breast-cancer/>.

[2] World Cancer Research Fund 2018, "*Breast Cancer*", American Institute of Cancer Research, viewed 15 November 2019, <www.wrcf.org/dietandcancer/breastcancer>.

[3] American Cancer Society 2019, "*What is Breast Cancer*" American Cancer Society, viewed 16 November 2019, <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>.

[4] McElroy, J A.Newcomb, P A.Trentham-Dietz, Titus-Ernstoff, L Hampton, J M. Egan, K M.,"BreastCancer Risk Associated With Electromagnetic field Exposure From Computer Work Ascertained From Occupational HistoryData",17 th conference, ISEE, September,2005.

[5] AbienFred M.Agarap, "On Breast Cancer Detection:An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset',International Conference on Machine Learning, February 2–4, 2018, Phu Quoc Island, Viet Nam.

[6] H.Asri, Mousannif, Hajar, Al Moatassime, Hassan, Noël, Thomas, "*Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis",* Procedia Computer Science,2016, vol. 83, pp.1064-1069.

[7] David A. Omondiagbe , Shanmugam Veeramani ,Amandeep S. Sidhu ,"*Machine Learning Classification Techniques for Breast Cancer Diagnosis"*, IOP Conference Series: Materials and Science,2019.

[8] M . Kumari, V. Singh,"Breast Cancer Prediction Systems", Procedia Computer Science, 2018, vol.132,pp. 371-376.

[9] K.Kourou, T.Exarchos , " Machine Learning Application in Cancer Prognosis and Diagnosis",Computational and Structural Biotechnology Journal,2014.
[10] J. Guo, M.Fung, F. Iqbal, Kuppen, R. Tollenaar,J. Lebran," Revealing Early Determinants of Occurance of Breast Cancer", Information Systems Frontiers,2017 issue 6, pp.1233-1241.
[11] Karabatak M, Ince MC," *An expert system for detection of breast cancer based on association rules and neural network",* Expert systems with Applications",2009 March,vol 36(2),pp.3465-3469.
[12] Kharya S, Soni S.*," Weighted naive bayes classifier: A predictive model for breast cancer detection*", International Journal of Computer Applications. 2016 Jan,vol.133(9), pp.32-37.
[13] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, *"A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning*,"Computer Structral Biotechnol.ogy,2017,vol. 15, pp. 75– 85.
[14] B.M.Gayathri.,C.P.Sumathi and T.Santhanam ," B*reast Cancer Diagnosis Using Machine Learning Algorithms –A Survey "*,International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013.
[15] PayamZarbakhsh,Abdoljalil Addeh*," Breast cancer tumor type recognition using graph feature selection technique and radial basis function neural network with optimal structure*", Journal of Cancer Research and Therapeutics,2018,vol.14,pp.625-33.
[16] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.
[17] Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M.Blondel, P. Prettenhofer, R. Weiss,V. Dubourg, J. Vanderplas, A. Passos, D. Cour-napeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011.,"*Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research* 12 (2011)",pp. 2825–2830.
[18] *F.Bunea,"Honest Variable Selection in Linear and Logistic Regression models*" International Journal of Statistics2 (2008).
[19] Octaviani TL, Rustam Z,"*Random forest for breast cancer prediction",*AIP Conference Proceedings, AIP Publishing Nov 4 ,2019 ,vol. 2168.
[20] LeCun Y, Bengio Y, Hinton," *Deep learning. nature.*", May,2015,issue 521(7553),pp.436-44.
[21] Medjahed SA, Saadi TA, Benyettou," *A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rule*s",International Journal of Computer Applications. 2013 Jan 1,vol.62(1).
[22] Yi L, Yi W," *Decision Tree Model in the Diagnosis of Breast Cancer"* ,International Conference on Computer Technology, Electronics and Communication (ICCTEC) , IEEE. Dec 19 ,2017 ,pp. 176-179 .
[23] Mert A, Kilic N, Akan A.,"*Breast cancer classification by using support vector machines with reduced dimension."*,ELMAR, ,IEEE,2011 Sep 14 ,pp.37-40.
[24] Gou J, Du L, Zhang Y, Xiong T ,"*A new distance-weighted k-nearest neighbor classifier",* J.ournal of Information of Computer Science,June 2012 ,vol.9(6),pp.1429-36.
[25] UCHealth 2015, "*How Accurate are mammograms?*", UCHealth viewed 16 November 2019, <www.uchealth.org/today/how-accurate-are-mammograms/>
[26] Kingma D P, Ba J, *"Adam: A method for stochastic optimization".* arXiv preprint arXiv:1412.6980. 2014 Dec 22.