A Machine Learning based Fine-Tuned and Stacked Model: Predictive Analysis on Cancer Dataset

Ravi Aavula
Guru Nanak Institutions Technical Campus,
Hyderabad, (T.S), India

R. Bhramaramba
GITAM Institute of Technology, GITAM,
Visakhapatnam, (A.P), India

***Abstract*—The earlier forecast and location of disease cells can be useful in curing the illness in medical applications. Knowledge discovery is having many significant roles in health sector, bioinformatics etc. Plenty of hidden information is available in the datasets present in the various domains like - medical information, textual analysis, image attributes exploration etc. Predictive analytics and modeling encompasses a variety of statistical methodologies from machine learning that can analyze the present along with historical facts to make the predictions about the future events. Breast cancer research already has involved with the good amount of progress in recent decade, but due to advancement in technologies, there is still some possibilities for an improvement.**

In this paper, the fine-tuned and stacked model procedure is presented which is experimented on standard breast cancer dataset. The obtained results show the improvement over state-of-the-art algorithms with improved performance parameters e.g. disease prediction accuracy, sensitivity and better F1 score etc.

***Keywords*—Machine learning; Cancer prediction; Data mining and Knowledge discovery; Supervised learning; Neural Networks**

I. Introduction

Over past decades, the life-style of individuals has modified underneath the various factors i.e. changes in the nutrition thing, utilization of the artificial agents or varied technological enhancements. It incorporates a negative as well as positive side. On one aspect, life is easier; there are new opportunities to carry out totally diverse kind of human motions etc. On another aspect, various types of diseases have increased and that require new varieties of treatment phenomenons to enhance a high quality of health care and the all in one solutions for associated well-being. Breast cancer [6] is one in all the foremost common cancer in conjunction with respiratory organ and cartilaginous tube cancer, glandular cancer and carcinoma among others. The employment of information science and machine learning approaches [11][13] in medical fields proves to be prolific as such approaches are also thought of nice help within the higher cognitive process of medical practitioners. Conventional strategic movements are hardly tolerating while facing with the multivariate natured data. The earlier forecast and location of disease cells with better accuracy can be useful in curing the illness in various medical / healthcare applications.

A. Related Work

M. Bruijne given a machine learning approach and framework [1] for disease detection and diagnosis for medical and healthcare data. Sherafatian [2] in his work, given a concept, tree-based procedures which is indicated as the minimal length subset of miRNA for the disease diagnosis. H. Wang et al. [3] proposed an ensemble procedure based on support vector regression for the cancer disease diagnosis. C. M. Lynch et al., in their work [4], have proposed a procedure for lung cancer patient's predictive analysis and their survival scenario chances via the supervised ML methods. Sommen et al. [5] given a method for predictive features for early cancer detection scenario. Wassan et al. [6] surveyed the various machine learning methods in Bioinformatics domain. N. Khuriwal et al. [7] proposed a framework for breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. Ali et al. [8] compared two techniques i.e. SVM and Neural semantic circuit networks cancer contamination diagnosis. They did experiments with the variety of kernel functionalities for support vectors hyperplanes e.g. mpl (with 86% accuracy), radial basis function (89%), quadratic (88%) furthermore polynomial (approx 88%).

Shajahaan et al. [9] utilized "Wisconsin Breast Cancer Database" for contamination analysis within variety of learning rules like Random Forests (RF), C4.5 etc. Chaurasia et al. [10] employed WEKA environment with 10-folds cv procedure for the algorithms like - k-NN, Best First randomized nodes forests, SMO etc. Authors of [11] in their work, performed a comparative analysis among Radial Basis Function kernel circuitary, Multilayer Perceptron circuit nodes forest along with the canonical logistical marking regression algorithmic procedures. Abed et al. [12] contemplated a categorization method of hybrid nature for cancer infected cell's interpretation and analysis. Ivankov et al. [13] given an extensive comparison of some significant and practically used machine learning routines in the binate natured classification elucidation.

B. Research Contribution

This paper contributes as follows:

- First, the state of the art scenarios and developments, results in the probem domain are reviewed. The limitations in existing methods drove the development of proposed solution for prediction model on cancer dataset.
- Then the experiments are performed on Breast Cancer Wisconsin (Diagnostic) Data Set [14] and further the novelty of proposed methodology through comparison results of existing approaches is proved on the same dataset.

C. Organization of the Paper

The remaining paper is structured as - Section 2 elaborates about some significant definitions and preliminaries. The proposed idea and detailed procedure is given in section 3. Experimental results along with comparative performance analysis are given in section 4. Finally, conclusive summary is given in section 5.

II. Preliminaries and Definitions

This sections presents some definitions and preliminaries.

A. Neural Networks Architecture

The basic design and working functionality of this classifier is somewhat equal to the human brain (concluded in lot of neurological studies). In the point of view of structure, its flexible accordingly to the task user needs to perform I.e. high dimension diminishing, localization, regression, categorization etc. First, based on input training data and learning algorithm, the classifier model is trained. The structure consists of i/p layer, hidden (middle) layer and o/p layer.[15] Here, the training process is more time consuming and it can perform multiclass classification. The accuracy of this classifier sometimes degrades due to less effective preprocessing and presence of missing values etc.

B. Support Vector Regression

This falls under the category of supervised learning mechanism where training and testing both phases are present. This classifier is based on the phenomenon of n dimension planes and hyperplanes. The n value depends on the total number of classes and the features lies on the planes spaces as separate data points. Here the task is to find support vectors/hyperplanes, which separate the data points into various available categories.[16] The separating hyperplanes can be linear in nature or nonlinear depending upon input problem. Here, the functionality of chosen kernels, variance plays significant role in output performance parameters.

C. Fuzzy SVM

If the attributes values and classes in the data are not discrete in nature means they are continuous natured then fuzzy logic is utilised in the support vector procedure.[17, 18] Some data which includes noise elements into it, should also be processed through fuzziness logic.

D. Bayesian Classifiers

As the metadata collection, some statistical and probability computations are first associated here.[19] To find the certain correlations among features, Bayesian and naive independent procedures are utilised. This classifier was invented in 1950s and since then it is being used with lot of variations in terms of correlations scanning, imputed variable predictors, dynamic features induction, non-linearized learning etc.

III. Proposed Methodology

This section presents the core idea and detailed algorithmic procedure.

A. Core Idea

Here, the curse of dimensionality and curse of multivariate nature of data has been dealt. This research contribution tried to achieve a computationally efficient dimension reduction and further classification based disease prediction procedural framework with comparatively improved and high accuracy. The detailed procedures are given as Algorithmic frameworks 3.2.1 and 3.2.2 in below sub-section. In algorithm 3.2.1, dimension reduction is performed as a pre-processing phase to deal with the curse of dimensionality. The dataset is converted into lower dimensional space (projection of a feature space into a smaller subspace), which is aimed to get rid of overfitting and perform reduction in computational cost. Further in algorithm 3.2.2, fine-tuned neural network with stacking is applied for disease prediction analysis.

B. Procedure Steps

Algorithm 1 Algorithmic procedure 3.2.1

1: Compute $\rightarrow$ the mean vectors m_i (d-dimensional) for (i = 1, 2 $\cdots$, C_n); where, C_n: total no. of different classes for the dataset.

2: Compute $\rightarrow$ scatter matrices in two aspects -
Within-class matrix: Use eq. $\text{SMAT}_W = \sum_{i=1}^{c} S_i$; where,

$$S_i = \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T$$

$$m_i = \frac{1}{n_i} \sum_{x \in D_i}^{n} x_k$$

Between-class matrix: Use eq. $\text{SMAT}_B = \sum_{i=1}^{c} N_i(m_i - m)(m_i - m)^T$; where m, m_i and N_i are - overall mean, sample mean and sizes of respective categories.

3: Compute $\rightarrow$ Eigenvectors $(ev_1, ev_2, \cdots, ev_d)$ along with the corresponding Eigenvalues $(\lambda_1, \lambda_2, \cdots, \lambda_d)$ for scatter matrices computed in step 2.

4: Make the tuples list and perform sorting of Eigenvectors with decrement in Eigenvalues.

5: k Eigenvectors are chosen with the largest of Eigenvalues and construct the d $\times$ k dimension matrix $\mathcal{M}$ (here, each column denotes an eigenvector).

6: Transform samples into new subspace by using d $\times$ k matrix i.e. -

$$Q_{(n \times k)} = P_{(n \times d)} \times \mathcal{M}_{(d \times k)}$$

Here, P denotes: $n \times d$ sized matrix possessing n samples, Q denotes: transformed $n \times k$ sized samples in reduced new feature-subspace.

Further, neural networks with MLP (multi-layer perceptron) architecture is used with finetuning for classification based disease prediction. Neurons are seperated in form of input layer, hidden layer and output layer. Steps involved in the procedure are given as algorithm 3.2.2 -

Learning rate is a configuration parameter, used to control the amount of weights that need to be updated in model procedure. Steps 1-4 are forward neural propagation and steps 5-11 are backward neural propagation.

Algorithm 2 Algorithmic procedure 3.2.2

1: Consider, X: as i/p matrix and Y: as o/p matrix. Initialize $\rightarrow$ assign random values to weights and biases.
Consider, w_{hl}: weight matrix for hidden layer
b_{hl}: bias matrix for hidden layer
w_{ol}: weight matrix for output layer
b_{ol}: bias matrix for output layer

2: Perform linear transformation as -
$$Input_{hidden\_layer} = matrix\_dot\_product(X, w_{hl}) + b_{hl}$$

3: Using Activation function (sigmoid), perform the non-linear transformation -
$$Activation_{hidden\_layer} = sigmoid(hidden\_layer\_input)$$
sigmoid function will return output as = $\frac{1}{1+e^{-x}}$

4: Perform -
$$o/p\_layer\_input = matrix\_dot\_product(hiddenlayer\_activations \times w_{ol}) + b_{ol}$$
o/p = sigmoid(o/p_layer_input)

5: Prediction is compared with actual output and calculate gradient of error.
Error(E) = Y - output

6: Compute -
$$Slope_{o/p\_layer} = derivatives\_sigmoid(output)$$
$$Slope_{hidden\_layer} = derivatives\_sigmoid(hiddenlayer\_activations)$$

7: Compute delta ($\triangle$) at output layer -
$$\triangle_{output} = E \times Slope_{o/p\_layer}$$

8: Error back propagation is done as -
$$E_{hidden\_layer} = matrix\_dot\_product(\triangle_{output}, w_{ol}^T)$$

9: Compute -
$$\triangle_{hidden\_layer} = E_{hidden\_layer} \times Slope_{hidden\_layer}$$

10: Update weights in network -
$$w_{ol} = w_{ol} + matrix\_dot\_product((hiddenlayer\_activations)^T, \triangle_{output}) \times value_{learning\_rate}$$
$$w_{hl} = w_{hl} + matrix\_dot\_product(X^T, \triangle_{hidden\_layer}) \times value_{learning\_rate}$$

11: Biases updated as -
$$b_{hl} = b_{hl} + sum(\triangle_{hidden\_layer}, axis = 0) \times value_{learning\_rate}$$
$$b_{ol} = b_{ol} + sum(\triangle_{output}, axis = 0) \times value_{learning\_rate}$$

IV. Experimental Analysis

This section presents an extensive experimental analysis. First the experimental envoirnmental setup, input dataset and its properties are discussed. Next, experimental output results are given. Later, in comparison table, the results of proposed procedure with significant existing approaches are compared.

A. Experimental setup

System specifications (Software and Hardware) used are as follows - OS: Ubuntu 16.04 LTS, 64 bit is used; hardware consists 4 GB RAM size along with Intel core i3 4030U CPU processor @1.90GHz $\times$ 4 clock speed. R language environment is used in the experimental analysis, with the specifications - R version - 3.4.4; RStudio Version - 1.0.136. R is a programming language and software framework for statistical analysis and computer graphics.

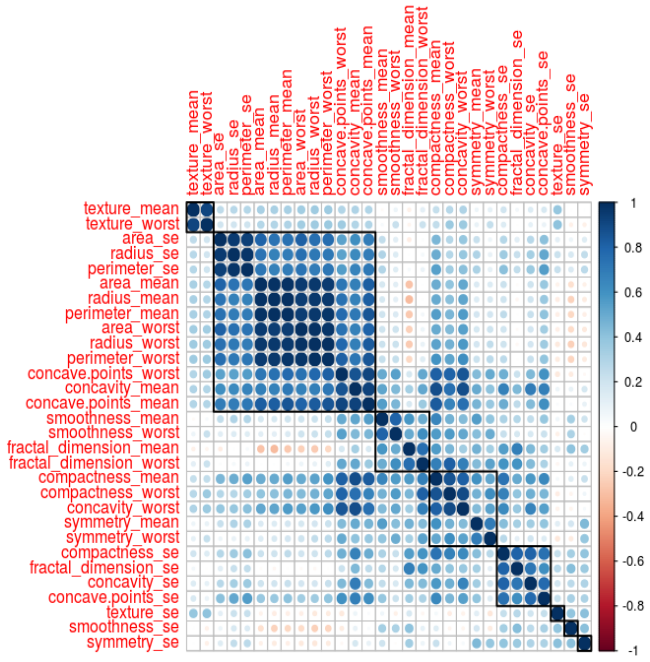

Fig. 1. Attributes correlation matrix

B. Input Dataset Details

In the experiments, *Breast Cancer Wisconsin (Diagnostic) Dataset* [14] is used. The details of this input dataset are as below:-

- Data Set Characteristics: Multivariate
- Attribute Characteristics: Real
- Number of Instances: 569
- Number of Attributes: 32
- Missing Values: N/A
- Ten real-valued features are computed for each cell nucleus
- Class distribution: 357 benign, 212 malignant.

C. Output Results Discussion

The experiments are performed on input dataset utilizing the proposed procedure. First the algorithmic procedure 3.2.1 is executed, and reduced dimensioned variables' sub-space is obtained. The attributes / variables correlation matrix is shown as Fig-1. Disease categorization for two classes of the dataset i.e. Benign(B) and Malignant(M) density distribution plot is given as Fig-2. Resultant confusion matrix after execution of the disease predictive analysis model is as follows -

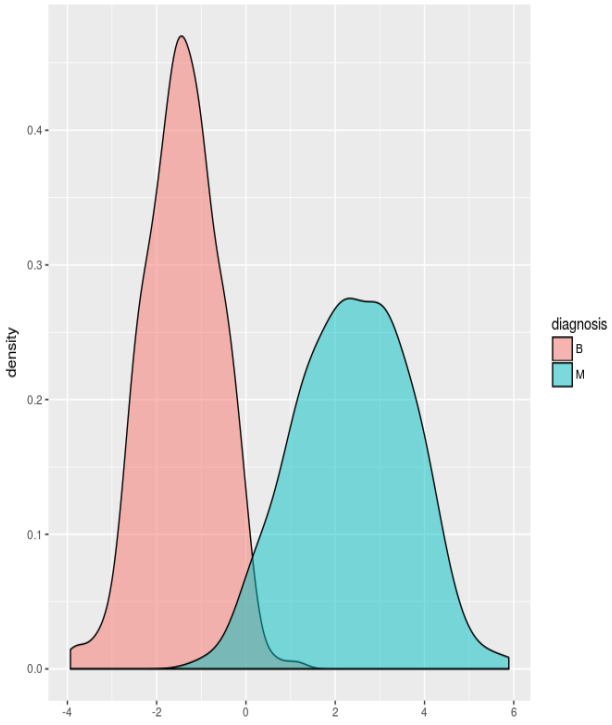

Fig. 2. Classes density distribution plot

TABLE I. EXPERIMENT RESULTS

Performance Parameters	Value
Accuracy	98.8%
P-Value	2e-16
Kappa coefficient	0.97
Sensitivity	0.98
Specificity	0.99

	Reference	
Pred	*Benign*	*Malignant*
Benign	106	1
Malignant	1	62

70:30 split ratio is used for the training and testing sample. Further, in the experimental procedure, it runs 5-folds cross validation (CV) on training set for cancer disease class prediction accuracy in test-set. Each experiment is repeated 5-times to verify the obtained accuracy of the generated models. Other statistical values / performance parameters obtained in the model (cancer disease predictive analysis model) are given in Table-1.

Graphical representation in shown as Fig-3. Better test accuracy as 98.8% is obtained along with other statistical performance parameters for disease prediction model performance i.e. P-value as 2e-16 (P-value represents, asymptotic significance for the model), Cohen's Kappa coefficient as 0.97, Sensitivity and Specificity as 0.98 and 0.99 respectively.

D. Comparative Performance Analysis

This section presents the comparative analysis where the obtained experimental results are compared with some significant existing state of the art methods on the same dataset. The comparison is given in Table-2. From comparative analysis, it is proved that the proposed procedure outperforms over other existing approaches.

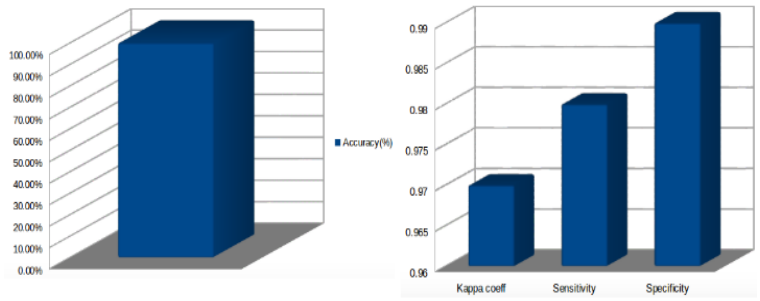

Fig. 3. Experimental analysis results: Performance parameters

TABLE II. COMPARATIVE ANALYSIS

Reference	Method	Dataset Split ratio	Prediction accuracy
E. Zafiropoulos et al. [20]	SVM with Gaussian RBF	70:30	89.28%
Ali et al. [8]	SVM with quadratic kernel	70:30	88%
Ali et al. [8]	SVM with poly kernel	70:30	88%
Ali et al. [8]	Neural Networks model	75:25	92%
S.S. Shajahaan et al. [9]	ID3 algorithm	-	92.99%
S.S. Shajahaan et al. [9]	C4.5 algorithm	-	95.57%
S.S. Shajahaan et al. [9]	CART	-	92.42%
S.S. Shajahaan et al. [9]	Naive Bayes	-	97.42%
J. Ivancakova et al. [13]	SVM	70:30	97.66%
J. Ivancakova et al. [13]	Random Forests	70:30	97.37%
J. Ivancakova et al. [13]	C4.5	70:30	95.61%
Proposed approach	fine-tuned and stacked DR-NN model	70:30	98.8%

V. CONCLUSIVE SUMMARY

Efficient preprocessing mechanisms to make the intelligent learning and prediction systems capable of dealing with multivariate data and effective learning technologies to find out the rules to describe the data are still of urgent need. Some limitations in the existing methods motivated us to propose a machine learning based fine-tuned and stacked model, utilizing which the experimental analysis is performed on cancer dataset in an computationally efficient manner with improved disease prediction accuracy.

REFERENCES

[1] Marleende Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, Medical Image Analysis, Volume 33, October 2016, Pages 94-97.

[2] Masih Sherafatian. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping, Gene, Volume 677, 30 November 2018, Pages 111-118.

[3] Haifeng Wang, BichenZheng, Sang Won Yoon, Hoo Sang Ko. A support vector machine-based ensemble algorithm for breast cancer diagnosis, European journal of Operational Research, Volume 267, Issue 2, 1 June 2018, Pages 687-699.

[4] Chip M.Lynch, Behnaz Abdollahi, Joshua D.Fuquac, Alexandra R.de Carlo, James A.Bartholomai, Rayeanne N.Balgemann, Victor H.van Berkel, Hermann B.Frieboes. Prediction of lung cancer patient survival via supervised machine learning classification techniques, International Journal of Medical Informatics, Volume 108, December 2017, Pages 1-8.

[5] Fonsvan der Sommen, Sander R.Klomp, Anne-FrSwager, Svitlana Zinger, Wouter L.Curvers, Jacques J.G.H.M.Bergman, Erik J.Schoon, Peter H.N. de. Predictive features for early cancer detection in Barrett's esophagus using Volumetric Laser Endomicroscopy, Computerized Medical Imaging and Graphics, Volume 67, July 2018, Pages 9-20.

[6] Jyotsna T. Wassan, Haiying Wang, Huiru Zheng. Machine Learning in Bioinformatics, Reference Module in Life Sciences, 2018.

[7] Naresh Khuriwal, Nidhi Mishra. Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm, IEEMA Engineer Infinite Conference (eTechNxT), IEEE, 13-14 March 2018.

[8] E. E. E. Ali, W. Z. Feng. Breast Cancer Classification using Support Vector Machine and Neural Network, International Journal of Science and Research, vol. 5, issue 3, 2016, pp. 1-6.

[9] S. S. Shajahaan, S. Shanthi, V. ManoChitra. Application of Data Mining Techniques to Model Breast Cancer Data, International Journal of Emerging Technology and Advanced Engineering. vol. 3, issue 11, 2013, pp. 362-369.

[10] V. Chaurasia, S. Pal. A Novel Approach for Breast Cancer Detection using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, issue 1, 2014, pp. 2456-2465.

[11] J. Padmavathi. A Comparative study on Breast Cancer Prediction Using RBF and MLP, International Journal of Scientific & Engineering Research, vol. 2, issue 1, 2011, pp. 1-5.

[12] B. M. Abed, K. Shaker, Hamid A. Jalab. A hybrid classification algorithm approach for breast cancer diagnosis, Proceedings of IEEE Industrial Electronics and Applications Conference (IEACon 2016), Malaysia, 2016, pp. 264-269.

[13] Juliana Ivancakova, Frantiek Babi, Peter Butka. Comparison of Different Machine Learning Methods on Wisconsin Dataset, IEEE 16th World Symposium on Applied Machine Intelligence and Informatics, Slovakia, February 7-10, 2018.

[14] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+ Wisconsin+%28 Diagnostic%29

[15] S Agatonovic-Kustrin, R Beresford. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, Journal of Pharmaceutical and Biomedical Analysis, Volume 22, Issue 5, June (2000), Pages 717-727.

[16] Cortes, C., Vapnik, V. (1995). Support-vector networks, Machine Learning, 20(3), 273-297.

[17] Abonyi, J., Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers, Pattern Recognition Letters, 24(14), 2195-2207.

[18] C.-F. Lin and S.-D. Wang, Fuzzy support vector machines, IEEE Transactions on Neural Networks, vol. 13, no. 2, March 2002.

[19] Nir Friedman, Ron Kohavi. Bayesian classification, Stanford Artificial Intelligence Laboratory, (1999).

[20] E. Zafiropoulos, I. Maglogiannis, and I. Anagnostopoulos. 2006. A support vector machine approach to breast cancer diagnosis and prognosis, Artificial Intelligence Applications and Innovations (2006), 500-507.