Different Perspective of Machine Learning Technique to Better Predict Breast Cancer Survival

Kaushal Kumar
Institute of Applied Mathematics
Heidelberg University
Germany, 69117 Heidelberg
Email: kaushal.kumar@stud.uni-heidelberg.de

Vijay Vikram Singh
SCIS, Jawaharlal Nehru University
New Delhi-110067, India
Email: vvsvikramsingh@gmail.com

Ramakrishna Ramaswamy
Department of Chemistry
Indian Institute of Technology Delhi
New Delhi-110016, India
Email: ramaswamy@iitd.ac.in

***Abstract*—Machine learning plays a key role in the aid of cancer diagnosis and detection. The researcher implements different techniques of machine learning to check and detection of breast cancer. Some researchers claim that their algorithms are more accurate, faster, and easier than others. My study relies on recently developed machine learning algorithms like genetic algorithms and deep belief nets. I' m interested to construct a system to accurately differentiate between benign and malignant breast tumors. We've optimized the training algorithm. During this context, we applied the genetic programming technique to decide on the only real features and ideal parameter values of the machine learning classifiers. The studies depend on sensitivity, cross-validation, accuracy, and ROC curves Among all the differing types of classifier utilized in this paper genetic programming is the foremost effective model for feature selection and classifier.**

***Index Terms*—Breast cancer, Deep Learning, Deep belief nets, Genetic Algorithm.**

I. Introduction

The study of cancer research isn't new using Machine learning. Artificial neural networks (ANNs) and decision trees (DTs) are employed in cancer detection and diagnosis for nearly 20 years. Today the role of machine learning methods has utilized in an exceedingly big selection of applications starting from detecting and classifying tumors via X-ray and CRT images to the classification of malignancies from proteomic and genomic (microarray) assays [1]. In line with the foremost recent PubMed statistics, over 1600 papers are published on the subject of machine learning and cancer [2]. these papers have the interest to identify, classify, detect, or distinguish tumors and other malignancies using machine learning methods. In other words, machine learning plays a key role in aid of cancer diagnosis and detection [3].

Breast cancer could also be a prevalent explanation for death, widespread among women worldwide [4]. Different techniques are developed for early detection and treatment of breast cancer and reduced the quantity of deaths, and different aided breast cancer diagnosis methods are aware of increase the diagnostic accuracy [5]. within the previous number of decades, Machine learning techniques which require three main stages: preprocessing, feature extraction, and classification. During this process, singular value decomposition (SVD) or principal component analysis (PCA) is applied to scale back the dimensionality of the feature representation. Many works have attempted to automate the diagnosis of breast cancer supported machine learning algorithms [6].

Nowadays, machine learning is playing a very important role within the medical sciences. Unfortunately, machine learning continues to be a field with high barriers and sometimes requires expert knowledge. Designing an honest machine learning model including the stages of preprocessing, feature selection, and classification processes requires a bunch of skills and expertise. The expert in machine learning chooses the acceptable technique for this problem domain. However, the nonexperts in machine learning spend many some time to optimize their proposed models and to attain the target performance. during this context, the aim of the work is to automate the design of the machine learning models employing dozens of techniques. and Genetic programming is that the best combination to optimized the techniques.

There are two challenges to automate the breast cancer diagnosis: (i) to find out which model most closely fits the information and (ii) a way to automatically design machine learning model by adjusting the parameters.

In Section 2,3,4 we have explained material methods and mathematical details of algorithms. Section 5 explains the computational results, whereas the foremost conclusions is discussed in Section 6.

II. Materials and Methods

Dataset Used for research: We usethe breast cancer, Coimbra, Data Set downloaded from the UCI Machine learning Repository. This dataset [7] employed by Patricio to predict cancerous and noncancerous tumors. Blood sample was all collected at the identical time of the day after overnight fasting. Clinical, demographic, and anthropocentric data were collected for all participants, under similar conditions, always by the identical research physician during consultation. Collected data included age, weight, height, and menopausal status (for each participant, this status expressed whether she

was a minimum of 12 months after menopause or reported a bilateral oophorectomize). During this process, clinical features, including age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP-1 were observed or measured for every of the 166 patients. There are 10 predictors, all quantitative, and a binary quantity, indicating the presence or absence of breast cancer. The predictors are anthropocentric data and parameters which may somewhat be gathered in routine blood analysis [8].

Background: For analysis purpose we use several machine learning techniques like decision tree, random forest naive bayes, Linear SVM, Logistical regression and KNN to get possible biomarkers.

Feature Selection: In this process we try to find minimum number of effective features. Feature selection methods could even be divided into the wrapper, filter, and embedded methods [9]. During this text, we make the interested reader alert to the probabilities of feature selection, providing a basic taxonomy of feature selection techniques, and discussing their use, variety, and potential during diffusion of both common similarly as upcoming bioinformatics applications [10].

A. Genetic Algorithm

In 1975 Holland developed the Genetic Algorithm [11]. In his book, Adaptation in Natural and Artificial Systems said the Genetic Algorithm could even be a computational optimization paradigm modeled on the concept of biological evolution. The advantage of a random search to the goal of the foremost effective subset. Genetic Algorithm is an optimization procedure that operates in binary search spaces and manipulates a population of potential solutions. It evaluates fitness function by a get above-average performance with high performance.
Genetic algorithm system encodes the solution into a numerical form called a chromosome that consists of the smallest genetic component of the gene [12]. Then, randomly selected to form a spread of chromosomes thanks to the initial population. Then there will be selected pairs of chromosomes from the population for crossover and mutation operations, repeated until the result stops within the desired condition that the utmost number of generations is achieved. After that, the fitness values of each chromosome are calculated. the foremost effective possible fitness value of the chromosome is visiting be selected within the gene column for next-generation reproduction [13]. The algorithms on this method could even be described as follows:

Step 1: Generating initial population.
Step 2: Evaluate the fitness value.
Step 3: Selection process from the initial population to conduct crossover operations.
Step 4: Crossover of each pair of selected parent chromosomes.
Step 5: Mutation on the chosen chromosome, and calculate the fitness value of the chromosome.
Step 6: After mutation,evaluated the fitness value the newly generated child population. If the fitness value of the offspring is higher, then the parent chromosome is replaced with a replacement offspring chromosome. If not, then don't go spinning.
Step 7: If its fail repeats the step (3) to (6) until it reaches the foremost number of generations.

B. Classification Using Support Vector Machines

In 1998 Vapnik developed the Support Vector Machine (SVM). It constructs a classifier which divides training samples. And also, maximizes the minimum distance or margin [14]. Consider the set of two classes, which can somewhat be described as

$$(t_i, t_j), \qquad i = 1, 2, ..., n. \qquad t \in R^n, x \in \{-1, +1\}$$

where x_i refers to class labels and i denotes the number of data. The basic idea of this method is to construct the hyperplane that separated the two classes of data with the equation

$$x(t) = w^T.t + b$$

where b is that the bias and w is vector perpendicular (n-dimensional) to hyperplane. The maximize of margin is like minimizing the load norm $\| w \|^2$ defined to

$$x_i(w.t + b) \geq 1, \qquad i = 1, 2, ..., n$$

By introducing positive Lagrange multipliers $\alpha_i, i = 1, 2, ..., n$ for the above constraints, formed a primal Lagrange function as follows

$$L_P = \frac{1}{2} \| w \|^2 - \Sigma_{i=1}^n \alpha_i [x_i(w.t + b) - 1]$$

That is a convex quadratic programming problem to resolve the "dual" problem. Its means maximize L_p subject to the ∇L_p with relevancy w and b vanishes. the solution of dual problem's

$$\alpha = argmin \frac{1}{2} \Sigma_{i=1}^n \Sigma_{j=1}^n \alpha_i \alpha_j x_i x_j (t_i, t_j) - \Sigma_{i=1}^n \alpha_i$$

is given by above equation with constraints

$$\alpha_i \geq 0, \qquad i = 1, 2, ..., n$$
$$\Sigma_{i=1}^n \alpha_i x_i = 0$$

Then the solution is these samples with nonzero α_i are called support vectors, which are critical samples for classification [15]. we'll map the training data to higher dimensional space to create a nonlinear decision surface. The separating hyperplane to maximize the margin and hyperplanes is represented to kernel function functions as follows

$$H(t_i, t_j) = \psi(t_i).\psi(t_j)$$

The decision function is defined in formula these kernels as [16]

$$f(t) = sign(\Sigma_{i=1}^n \lambda_i x_i H(t_i.t_j)) + b$$

where λ_i are coefficients of Lagrange multipliers.

III. Restricted Boltzmann Machine

Restricted Boltzmann Machine (RBM) [17] could be a two-layer undirected graphical model within which the first layer consists of visible input variables $\mathbf{x} \in \{0,1\}^D$, and a second layer consists of hidden variables (units) $\mathbf{h} \in \{0,1\}^M$. We allow only the inter-layer connections, i.e., there aren't any connections within layers. Moreover, we add a third layer that represents observable output variable $y \in \{1,2,...,K\}$. we use the 1-to-K coding scheme which finally ends up in representing the output as a binary vector of length K denoted by y, specified if the output (or class) is k, then all elements are zero except element y_k which takes the worth 1.

A RBM with M hidden units might be a parametric model of the joint distribution of visible and hidden variables, that takes the form:

$$p(\mathbf{x},\mathbf{y},\mathbf{h}|\theta) = \frac{1}{Z(\theta)} e^{-E(\mathbf{x},\mathbf{y},\mathbf{h}|\theta)}$$

with parameter $\theta = \{\mathbf{b},\mathbf{c},\mathbf{d},\mathbf{W}^1,\mathbf{W}^2\}$, and where:

$$E(\mathbf{x},\mathbf{y},\mathbf{h}|\theta) = -\mathbf{b}^T\mathbf{x} - \mathbf{c}^T\mathbf{h} - \mathbf{d}^T\mathbf{y} - \mathbf{x}^T\mathbf{W}^1\mathbf{h} - \mathbf{h}^T\mathbf{W}^2\mathbf{y}$$

is the energy function, and

$$Z(\theta) = \Sigma_{\mathbf{x},\mathbf{y},\mathbf{h}} E(\mathbf{x},\mathbf{y},\mathbf{h}|\theta)$$

is a partition function.

This model is termed Restricted Boltzmann Machine (RBM) [18]. The condition probability of a configuration of the visible unit **,** given a configure of the hidden units **h**, is

$$p(\mathbf{x}|\mathbf{h}) = \Pi_i p(x_1|\mathbf{h})$$

$$p(x_i = 1|\mathbf{h}) = \sigma(b_i + \mathbf{W}_i^1.\mathbf{h})$$

$$p(y|\mathbf{h}) = \frac{e^{d_y} + (\mathbf{W}_{.y}^2)^T\mathbf{h}}{\Sigma_{\bar{y}} e^{d_y} + (\mathbf{W}_{.y}^2)^T\mathbf{h}}$$

$$p(\mathbf{h}|y,\mathbf{x}) = \Pi_j p(h_j|\mathbf{h}$$

$$p(h_j = 1|y,\mathbf{x}) = \Pi_j \sigma(c_j + W_{jy}^2 + (\mathbf{W}_j^1)^T\mathbf{x})$$

where σ is that the logistic sigmoid function, $\mathbf{W}_i^l$ is that the ith row of weight matrix $\mathbf{W}^l$, $\mathbf{W}_j^l$ is that the jth column of weight matrix $\mathbf{W}^l$, W_{ij}^l is that the element of weights matrix $\mathbf{W}^l$.

It is possible to exactly compute the distribution $p(y|\mathbf{x}$ which can be further accustomed choose the foremost probable class label. This conditional distribution takes the following form

$$p(y|\mathbf{x}) = \frac{e^{d_y}\Pi_j(1 + (e^{c_j + (\mathbf{W}_j^1)^T\mathbf{x}})e^{W_{jy}^2})}{\Sigma_{\bar{y}} e^{d_{\bar{y}}}\Pi_j(1 + (e^{c_j + (\mathbf{W}_j^1)^T\mathbf{x}})e^{W_{j\bar{y}}^2})}$$

Pre-computing the terms $c_j + (\mathbf{W}_j^1)^T\mathbf{x}$ allows to chop back the time needed for computing the conditional distribution to $O(MD + MK)$. within the medical context, this chance expresses the probability of occurring ith input for the given other inputs and class labels.

Typically, the parameter θ in RBM are learned from the knowledge using the likelihood function:

$$p(\mathbf{x},\mathbf{y},|\theta) = \frac{1}{Z(\theta)} \Sigma_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{y},\mathbf{h}|\theta)}$$

There are different inductive principle for learning RBM, e.g., Approximate Maximum Likelihood, Maximum Pseudo-Likelihood, ratio Matching [19]. The widely-used methos is Constrastive Divergence, which is used during this paper.

IV. Deep belief networks (DBN)

DBN is fine-tuned as neural networks, DBN has many non-linear hidden layers, DBN is generatively pre-trained beside it can act as non-linear dimensionality reduction for input features vector, and eventually, the networked teacher is another sensory input [20]. Let the output neurons state y represent the training example y^* .The discrepancy between the expected output y^* and thus the actual output y is using the squared error measure:

$$E = (y^* - y)^2$$

The change in weight, which is added to the old weight, is adequate the merchandise of the training rate and also the gradient of the error function, multiplied by 1:

$$\triangle w_{ij} = -\frac{\partial E}{\partial w_{ij}}$$

where the majority data is unlabeled. However, backpropagation neural network requires a labeled training data [22], [23]. In recent years, attention has shifted to deep learning [24]. Deep learning [21] would be a gaggle of algorithms in machine learning that attempts to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations [25]. RBM may be a stochastic artificial neural network which learns a probability distribution over its set of inputs, whereas, DBN could be a graphical model composed of multiple layers of latent variables (hidden units), with connections between the layers, but not between units within each layer [26].

From Hintons perspective, the DBN is viewed as a composition of easy learning modules each of which can be a restricted quite RBM that contains a layer of visible units. This layer represents knowledge. Another layer of hidden units represent features that capture higher-order correlations within the knowledge. the 2 layers are connected by a matrix of symmetrically weighted connections (W) and there don 't seem to be any connections within a layer [27].

The key idea behind DBN is its weight (w), learned by an RBM define both $p(v|h,w)$ then the prior distribution over hidden vectors $p(h|w)$ [27]. The probability of generating a visual vector is written as

$$p(v) = \Sigma_h(p(h|w)p(v|h,w))$$

As the learning of DBN is additionally a computationally intensive task, Hinton showed that RBMs could even be stacked and trained in an exceedingly very greedy manner to create the DBN [28]. He introduced a quick algorithm for learning DBN [28]. The load update between visible v and hidden h unites simply as:

$$\triangle w_{ij} = \varepsilon(\langle v_i, h_j \rangle^0 - \langle v_i, h_i \rangle^1)$$

where 0 and 1 within the above equation designate to the network data and reconstruction state, respectively.

V. Experimental Results

We used Python 3.7.4 and import various libraries for data handling so as that it can read datasets, then we've imported other libraries for data illustrate so as that we are able to easily understand the data sets and interpret furthermore. We've identified NaN values throughout the row likewise as a column so as that we are able to fill it with various methods sort of a global mean or with the mean of row or column etc [29]. For organization, interpretation, presentation, and an intensive understanding of the information. We've described the knowledge set, so we could know the mean, variance, minimum, maximum limit, and thus, the precise distribution of knowledge set [30]. As we all know that ultimately our data is assessed into binary classification which we've got illustrated the classification into a bar chart.

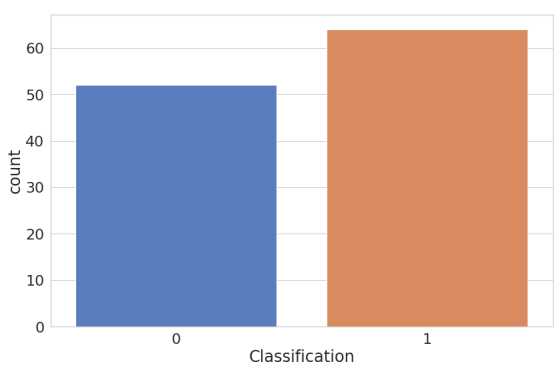


Fig.1 : Classification labels

Descriptive statistics were calculated for all variables employed in the study using python syntax. we've got identified the assorted arrangement with relation to binary classification, which creates us know the mean, median, mode, and variance of a specific classification. As we respond to through observation, each data carries with it nine entry and to grasp the distribution of feature in two classes, we've plated two or more plots in one figure so, that we are able to easily visualize each feature.

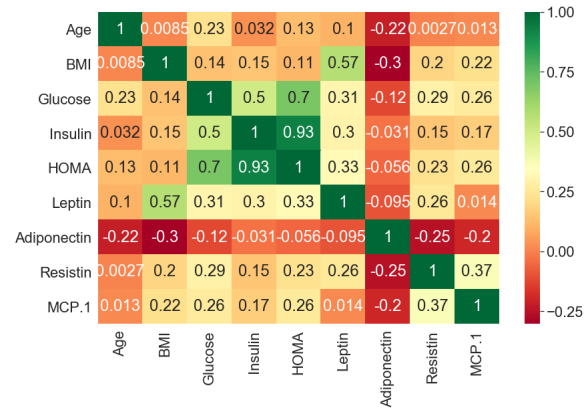


Fig.2 : Heat Map/Correlation plot

Heat Map : We've got also used another technique to simplify the complex data set i.e. heat map [31]. A heat map could be a two-dimensional representation of knowledge within which values are represented by color. There are often some ways to display heat maps, but all of them share one thing in common they use color to speak relationships between data values that will be much harder to know if presented numerically in an exceedingly spreadsheet.

A. Machine learning algorithms

We have used several supervised and unsupervised machine learning techniques to search out of the possible biomarker for breast cancer. to create a model for implementing the machine learning technique, Firstly we imported the library for the machine learning technique then we segregate data into tests and train to create a model. during this case, we've selected 25% data for testing and rest others for training.
Following Machine Learning techniques we've used with the subsequent accuracy:

Machine Learning Technique	Accuracy
Genetic Alogrithm	0.875000
Logistic Regression	0.862069
Linear SVM	0.827586
Random Forest	0.793103
Deep Belief Network	0.758621
Decision Tree	0.724138
Naive Bayes	0.724138
Radial SVM	0.551724
KNN	0.551724

Table: 1

As we see the above Machine learning technique with their accuracy we found the Fluctuation within the accuracy of models is because of lack of knowledge. we've only 116 individuals data and to organize a decent model we want to data a minimum of in a thousand. The experimental data on cancer is extremely scarce. There are many various methods are available for solving these problems, but lack of knowledge is clearly not an answer. a technique to beat this problem is to either we've to rearrange the important data.

B. K-Fold cross-validation

To overcome the matter with splitting into training and testing data sets, we've got applied K-fold cross-validation [32]. to use a machine learning technique, we've to specify what fraction of knowledge is employed for training and testing data sets and it'd be possible that some information in training data sets to urge the simplest results and a maximum number of knowledge points tested to urge the simplest validation [16]. Every information we've taken out of the training set into the test is lost for training sets. So, we've got to reset this swap. this can be where cross-validation comes into the image because this cause maximizes both of the sets cuz we have a good deal of information within the training sets to urge the most effective learning results, and also the maximum number of knowledge point we tested to induce to urge the simplest validation. the fundamental idea of cross-validation is that partition of knowledge set into k bin of equal size , for example: if we've got 300 datums and 10 bin size which means per bin have 30 data points, so, we've 30 data point in each bin. So in cross-validation, one among the k subsets, as testing test and remaining k-1 bin, are put together as training sets then train machine learning algorithms and test the performance on the testing set [33]. Machine learning algorithms run multiple time, during this case, 10 times and so the common the ten different testing sets performances for the ten different sets and average the test results for those k experiments . this can be the way by which we will use all data sets for training , and every one data for testing.

Machine Learning Technique	Cross-validation mean
Decision Tree	0.581818
Random Forest	0.562879
Naive Bayes	0.547727
Linear SVM	0.533333
Logistic Regression	0.509091
Radial SVM	0.509091
KNN	0.506818

Table: 2

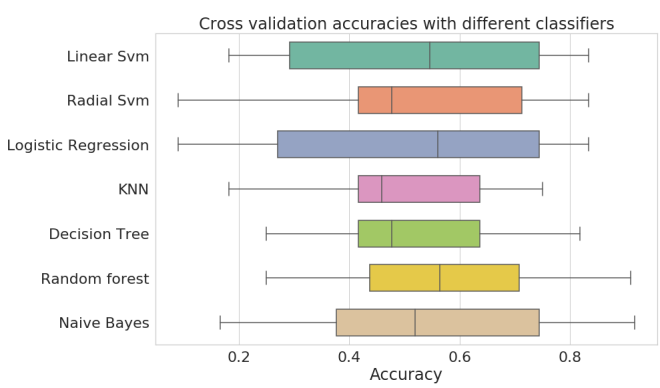


Fig.3 : Cross Validation accuracies with different classifiers

C. t-SNE Visualization

As we all know that we board a 3-D world in order that we are able to easily visualize the pattern within the 3-D, but what if we've data that is more complex? The visualization of knowledge is crucial steps to grasp the info and to grasp the complex data, dimension reduction is very important to step to require. Nowadays high dimensional data is everywhere and t-SNE [34] may be a popular technique for dimension reduction that we will use via sci-kit learn.

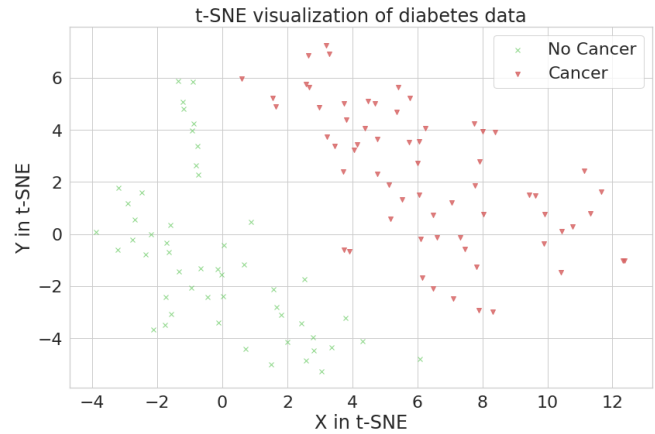


Fig.4 : t-SNE visualization of Cancer data

The complete list of parameters which are important in predicting breast cancer are as follows

Feature	Importance	Normalized	Cumulative
Resistin	33.4	0.213965	0.213965
Age	32.7	0.209481	0.423447
Glucose	31.9	0.204356	0.627803
BMI	16.9	0.108264	0.736067
HOMA	12.9	0.082639	0.818706
Leptin	10.5	0.067265	0.885971
Adiponectin	9.1	0.058296	0.944266
Insulin	4.5	0.028828	0.973094
MCP.1	4.2	0.026906	1.000000

Table: 3

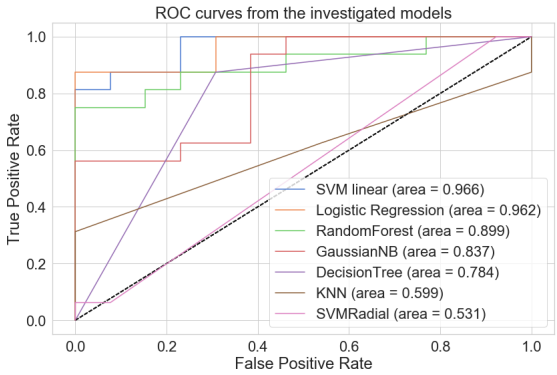
D. ROC curve

We have tested numbers of models to predict the possible outcomes and to understand what is the simplest model overall without having any quite hypothesis regarding the acceptance the speed of the model. one all told the foremost credit risk models relies upon sensitivity and specificity [35] which has evaluated by the receiver operating graph i.e. ROC curve [36]. In this, a graph is plotted by plotting the sensitivity against 1-specificity for every cutoff, the plot always starts with the lower-left corner where sensitivity is 0 and specificity is 1, in upper right corner correspond cut-off up to zero where sensitivity is 1 and specificity is 0

In general, speaking maximum area of the best left corner cover is healthier i.e. those curve cover maximum area of the best left have the next specificity associated with higher sensitivity [35]. The diagonal line divided the ROC curve into two halves, a degree above the diagonal i.e. upper left space represents good classification results contrariwise. Following formula for verity positive and false negative of ROC curve.

- $Sensitivity = \frac{TruePositive}{TruePositive+FalseNegative}$

- $Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$



VI. Conclusion

In this research, we've implemented different techniques of a Machine learning algorithm for detecting breast cancer supported deep belief nets, different type of regression and genetic algorithms. The overall neural network accuracy is 75.85% for DBN in breast cancer cases. Genetic algorithms give the upper accuracy of 87.50%. We believe that the proposed system is very helpful to the physicians for his or her final decisions on their patients. Since the info sets utilized during this work is little, if we've an oversized amount of information sets. These two methods used for prediction may be a more accurate prediction. By using such a tool, they'll make very accurate decisions.

Acknowledgment

KK thanks the India Department of Biotechnology (DBT): Joint DBT-Heidelberg University Graduate program on Big Data Research (fund: 2302086, assignment: 7815564) for support and RR is supported by the DST (Govt of India) through the JC Bose fellowship.

References

[1] A. L. D. Brown, J. Martin, "The role of apoptosis in cancer development and treatment response," *Nature Reviews Cancer*, 2005.

[2] S. Agrawal and J. Agrawal, "Neural network techniques for cancer prediction: A survey," *Procedia Computer Science*, vol. 60, pp. 769 – 774, 2015.

[3] K. D. Cole, H.-J. He, and L. Wang, "Breast cancer biomarker measurements and standards," *PROTEOMICS Clinical Applications*, vol. 7, no. 1-2, pp. 17–29, 2013.

[4] M. LLC, "MS Windows NT kernel description," 2018.

[5] M. Bacac and I. Stamenkovic, "Metastatic cancer cell," *Annual Review of Pathology: Mechanisms of Disease*, vol. 3, no. 1, pp. 221–247, 2008. PMID: 18233952.

[6] J. Malek, S. Abderrahim, S. Mabrouk, K. Torki, and R. Tourki, "Automated breast cancer diagnosis based on gvf-snake segmentation, wavelet features extraction and fuzzy classification," *Signal Processing Systems*, vol. 55, pp. 49–66, 04 2009.

[7] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using resistin, glucose, age and bmi to predict the presence of breast cancer," *BMC Cancer*, vol. 18, p. 29, Jan 2018.

[8] D. S.-S. A. L. G. M. G. M. P. L. L. A. B. S.-R. L. S. . R. S. Joana Crisstomo, Paulo Matafome

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000.

[10] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3240 – 3247, 2009.

[11] J. H. Holland, "Genetic algorithms and adaptation," in *Adaptive Control of Ill-Defined Systems*, pp. 317–333, Springer, 1984.

[12] A. M. E. M.-A. E. W. F. N. M. Dhahri, Habib, "Automated breast cancer diagnosis based on machine learning algorithms," *Journal of Healthcare Engineering*, 11 2019.

[13] K. Sastry, D. Goldberg, and G. Kendall, "Genetic algorithms," in *Search methodologies*, pp. 97–125, Springer, 2005.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] K. Polat and S. Gne, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing*, vol. 17, no. 4, pp. 694 – 701, 2007.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[17] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*, pp. 791–798, 2007.

[18] R. Salakhutdinov and G. Hinton, "An efficient learning procedure for deep boltzmann machines," *Neural computation*, vol. 24, no. 8, pp. 1967–2006, 2012.

[19] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 643–669, 2012.

[20] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 201–208, 2010.

[21] S. S. Haykin, *Neural networks and learning machines*. Upper Saddle River, NJ: Pearson Education, third ed., 2009.

[22] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

[23] J. A. Suykens, *Advances in learning theory: methods, models, and applications*, vol. 190. IOS Press, 2003.

[24] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Systems with Applications*, vol. 46, pp. 139 – 144, 2016.

[25] J. Dheeba, N. A. Singh, and S. T. Selvi, "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach," *Journal of biomedical informatics*, vol. 49, pp. 45–52, 2014.

[26] G. E. Hinton, "Deep belief nets.."

[27] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009. revision #91189.

[28] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[29] albertbup, "A python implementation of deep belief networks built upon numpy and tensorflow with scikit-learn compatibility," 2017.

[30] "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064 – 1069, 2016.

[31] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[32] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection,"

[33] 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 Systems with Applications*, vol. 41, no. 4, Part 1, pp. 1476 – 1482, 2014.

[34] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[35] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8 – 17, 2015.

[36] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.