GRNN and PNN Models for Cancer Prognosis and Prediction

Namrata Bilurkar[#1], **Ajay Prakash B V**[*2]
[#]Department of Information Science and Engineering,
SJB Institute of Technology, Kengeri, Bangalore, Karnataka, India

Abstract— **In this paper we are using Generalised Regression Neural Network (GRNN) and Probabilistic Neural Network model (PNN) to classify cancer patients into low and high risk groups. The data about the cancer patients has been obtained from SEER, a program of the National Cancer Institute which provides information on cancer incidence and survival in the United States. Application of Machine Learning (ML) techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BN), Support Vector Machines (SVM) and Decision Trees (DT) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. The classification of the cancer patients has been made easier by making use of the ML algorithms and here we are making use of the GRNN and PNN to classify the patients into high risk or low risk groups. The result of the evaluation will show the performance of the ML algorithms to model the risk of cancer on the patient's outcome.**

Keywords: **Generalised Regression Neural Network, Probabilistic Neural Network, Machine Learning, Artificial Neural Networks, Bayesian Networks, Support Vector Machines, Decision Trees**

I. INTRODUCTION

ML, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference. Every learning process consists of two phases: (i) estimation of unknown dependencies in a system from a given dataset and (ii) use of estimated dependencies to predict new outputs of the system. ML has also been proven an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an n-dimensional space for a given set of biological samples, using different techniques and algorithms. There are two main common types of ML methods known as (i) supervised learning and (ii) unsupervised learning. In supervised learning a labelled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labelled examples are provided and there is no notion of the output during the learning process. As a result, it is up to the learning scheme/model to find patterns or discover the groups of the input data. In supervised learning this procedure can be thought as a classification problem. The task of classification refers to a learning process that categorizes the data into a set of finite classes. Two other common ML tasks are regression and clustering. In the case of regression problems, a learning function maps the data into a real-value variable. Subsequently, for each new sample the value of a predictive variable can be estimated, based on this process. Clustering is a common unsupervised task in which one tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share. Suppose for example that we have collected medical records relevant to breast cancer and we try to predict if a tumor is malignant or benign based on its size. The ML question would be referred to the estimation of the probability that the tumor is malignant or no (1 = Yes, 0=No). There are several ML techniques, including and not restricted to the following – Decision Tree Learning, Association Rule Learning, Artificial Neural Networks, Inductive Logic programming, Support Vector Machines, Clustering, Bayesian Networks, Reinforcement Learning, Representation Learning, Similarity and Metric Learning, Sparse Dictionary Learning, Genetic Algorithms. Artificial neural networks (ANN) have emerged as a result of simulation of biological nervous system, such as the brain, on a computer. On the other hand, biological neural networks are much more complicated than the mathematical models used for ANNs. ANN was founded by McCulloch and co-workers beginning in the early 1940s [1]. They built simple neural networks to model simple logic functions.

II. RELATED WORKS

In [2], a probability network based heart failure program was developed to assist physicians in reasoning about patients, which produced appropriate diagnoses about 90% of the time on the training set. Azuaje et al. [3] employed artificial neural networks (ANN) to recognize Poincare-plot-encoded heart rate variability patterns related to the risk of the coronary heart disease. Tkacz et al. [4] demonstrates how wavelet neural networks (WNN) can be applied for disease classification useful to diagnose coronary artery disease at different levels. For the diagnosis of congenital heart diseases, Reategui et al. [5] proposed a model by integrating case-based reasoning with neural networks. In [6] fuzzy reasoning optimized by genetic algorithm was used for the classification of myocardial heart disease.

Derisi et al. [7] published that the expression patterns of many previously uncharacterized genes provided clues to their possible functions. Eisen et al. [8] presented that clustering gene expression data grouped together efficiently genes of known similar function. Shamir [9] described some of the main algorithmic approaches to clustering gene expression data. Getz et al. [10] presented two-way clustering approach to gene microarray data analysis. There are many researchers to attempt to predict colon cancer using various machine learning methods and they show that prediction rate of colon cancer can be approximately 80 –90%. Sarkar et al. [11] presented a novel and simple method that exhaustively scanned microarray data for unambiguous gene expression patterns. T class is a corresponding program of a method that incorporates feature selection into Fisher's linear discriminate analysis for gene expression based on tumor classification [12]. Li et al. investigated two Bayesian classification algorithms incorporating feature selection and these algorithms were applied to the classification of gene expression data derived from DNA microarrays [13]. Li et al. studied to decide which and how many genes should be selected [14]. Guyon et al. Proposed a new method of gene selection using support vector machine based on recursive feature elimination (RFE) [15]. Xiong et al. Reported that using two or three genes, one could achieve more than 90% accuracy of classification in colon cancer, breast cancer, and leukemia [36]. There are some related works on EANNs that combine the advantages of the global search performed by evolutionary algorithms and local search of the learning algorithms (like BP) of ANN. Yao [17] proposed EANNs approach, EPNet based on Fogel’s evolutionary programming (EP) as evolutionary algorithm. EPNet emphasizes the evolution of ANN behaviors by EP and uses a number of techniques, such as partial training after each architectural mutation and node splitting, to maintain the behavioural link between parent and its o9spring e9ectively. EPNet also encourages parsimony of evolved ANNs by attempting di9erent mutations sequentially. That is, node or connection deletion is always attempted before addition. EPNet has shown good performance in error rate and size of ANN. Cho proposed a new approach of constructing multiple neural networks that used genetic algorithms with speciation to generate a population of accurate and diverse ANNs. Speciation in genetic algorithm creates di9erent species, each embodying a sub-solution, which means to create diverse solutions not the best one [18]. Experiments with the breast cancer data from UCI benchmark datasets show that the method can produce more speciated ANNs and improve the performance by combining only representative individuals [19]. Several combination methods are applied to combine speciated neural networks [20].

III. NEURAL NETWORK MODELS

A traditional artificial neural network based on back propagation algorithm has some limitations. At first, the architecture of the neural network is fixed and a designer needs much knowledge to determine it. Also, error function of the learning algorithm must have a derivative. Finally, it frequently gets stuck in local optima because it is based on gradient-based search without stochastic property. Evolutionary algorithm is a kind of search method based on biological facts and uses a population of multiple individuals. The combination of evolutionary algorithm and neural network can overcome these shortcomings

Artificial Neural Networks (ANN) is inspired by the early models of information processing by the brain. ANN is an information processing paradigm that is simulated by the biological nervous systems towards learning process and is configured for a specific application, such as pattern recognition or data classification. A novel structure of large number of highly interconnected processing elements (neurons) and its synaptic connections are the key elements of this paradigm. A main problem in statistics with applications in many areas is to estimate a function from some instance of input-output pairs with little or no knowledge of the form of the function. This form of problem is called function approximation, inductive learning, and nonparametric regression. In neural networks terms this can solved using supervised learning process. The function is learned from the instances which a teacher supplies. As neural networks are extremely fast and efficient, we have considered GRNN and PNN to estimate the software development effort.

A. Generalized Regression Neural Networks

Generalised Regression Neural Network is a type of supervised learning model based on radial basis function (RBF) which can be used for regression, classification and time series predictions. The GRNN architecture is as shown in figure1.

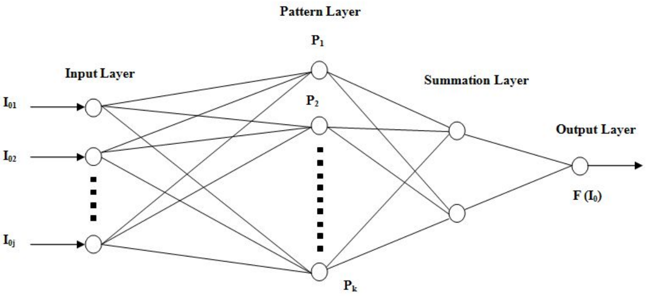

Fig. 1 Generalized Regression Neural Network (GRNN) Architecture

GRNN consists of four layers, which are named as input layer, pattern layer, summation layer and output layer. The number of input units depends on the total number of observation parameters i.e. an input vector ‘I’ (feature matrix F_i). The input layer connected to the pattern layer consists of neurons provides training patterns and its output to the summation layer to perform normalization of the resultant output set. Each of the pattern layers is connected to the summation neurons and calculates the weight vector using the following equations.

$$W_i = e^{\left[\frac{||I - I_t||^2}{2h^2}\right]}$$

$$F(I) = \frac{\sum_{i=1}^{n} T_i W_i}{\sum_{i=1}^{n} W_i}$$

Where the output F (I) is weighted average of the target values T_i of training cases I_i close to a given input case I.

B. Probabilistic Neural Network (PNN)

The PNN [19] is a Bayes–Parzen classifier. The foundation of the approach is well known decades ago (1960s). It models the Bayesian classifier & minimizes the risk of misclassification. Bayes’ classifier is usually criticized due to lack of information about the class probability distributions and makes use of nonparametric techniques, whereas the inherent advantage of PNN is the better generalization and convergence properties when compared to that of Bayesian classifier in classification problems. PNN Architecture is as shown in figure 2.

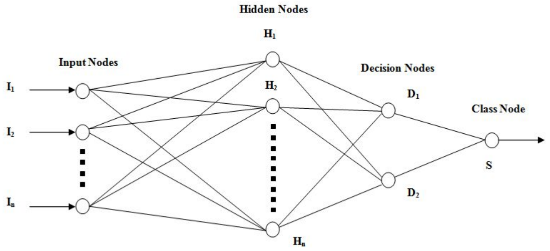

Fig. 2 Probabilistic Neural Network (PNN) Architecture

PNN is similar to that of supervised learning architecture, but PNN does not carry weights in its hidden layer. Each node of hidden layer acts as weights of an example vector. The hidden node activation is defined as the product of example vector 'E' &input feature vector 'F' given as h_i=E_i x F. The class output activations are carried out using the following equation

$$S_j = \frac{\sum_{i-1}^{n} e^{\frac{(h_i-1)}{\varphi^2}}}{N}$$

Where 'N' is example vectors belonging to class 'S', 'h_i' is hidden node activation and 'φ' is smoothing factor.

IV. PROPOSED METHODOLOGY

Proposed Methodology consists of data sets preparation, selecting the features from the data sets which are relevant, preparing training and test data sets. Apply the GRNN and PNN model

1. Data Collection of reported cancer patients in order to provide the training set to the supervised algorithms.
2. Feed the data sets to the algorithms.
3. Select the features to classify the test data.
4. Run the algorithms in order to obtain the MRE value.
5. Data set to classify the patients into their respective groups of risk levels.
6. Determine the accuracy levels on comparison of test results after running GRNN and PNN algorithms with the results obtained by running other ML algorithms.

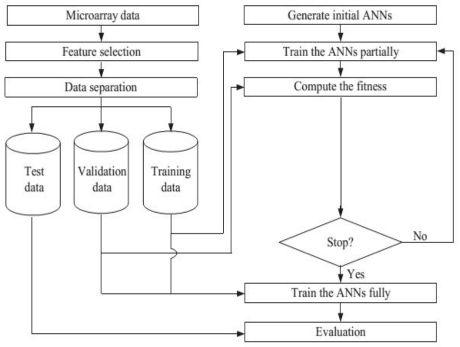

Fig. 2 Procedure for evaluating Neural Network Architecture

TABLE I
COMPARISION OF DIFFERENT METHOD WITH PROPOSED MODEL FOR CANCER PREDICTION

Author(s)	**Method**	**Type of data**	**Accuracy**
Park K et al.[21]	Graph-based SSL algorithm	SEER	71%
Delen D et al.[22]	DT	SEER	93%
Kim J et al.[23]	SSL Co-training algorithm	SEER	76%
Waddell M et al.[24]	SVM	SNPs	71%
Park C et al.[25]	Graph-based SSL Algorithm	Gene expression, PPIs	76.7%
Tseng C-J et al.[26]	SVM	Clinical, pathologic	68%
Proposed Model	**GRNN and PNN**	**SEER**	**94%**

Table 1 shows the comparison of different methods accuracy with proposed neural network model. Evaluation criterion is used to assess and the compare the performance of the GRNN and PNN models. Magnitude of Relative Error (MRE) and Mean Magnitude of Relative Error (MMRE) is used for evaluation of effort estimation.MRE is defined as in:

$$MRE = \frac{|\,ActualEffort - \Pr edictedEffort\,|}{ActualEffort} x100$$

MMRE for N projects is defined as in

$$MMRE = \frac{1}{N}\sum_{i=1}^{N} MRE_i$$

A higher value means worst prediction accuracy for MRE and MMRE,

V. CONCLUSION

Machine learning models can be used in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.

REFERENCES

[1] Haque ME, Sudhakar KV. ANN back propagation prediction model for fracture toughness in microalloy steel. Int J Fatique 2002;24:1003–10.

[2] Long, W. J., Naimi, S., & Criscitello, M. G. (1992). Development of a knowledge base for diagnostic reasoning in cardiology. Computers in Biomedical Research, 25, 292–311.

[3] Azuaje, F., Dubitzky, W., Lopes, P., Black, N., &Adamsom, K. (1999). Predicting coronary disease risk based on short-term RR interval measurements: A neural network approach. Artificial Intelligence in Medicine, 15, 275–297.

[4] Tkacz, E. J., & Kostka, P. (2000). An application of wavelet neural network for classification patients with coronary artery disease based on HRV analysis.Proceedings of the Annual International Conference on IEEE Engineering in Medicine and Biology , 1391–1393.

[5] Reategui, E. B., Campbell, J. A., & Leao, B. F. (1997). Combining a neural network with case-based reasoning in a diagnostic system. Artificial Intelligence in Medicine, 9, 5–27.

[6] Tsai, D. Y., & Watanabe, S. (1998). Method optimization of fuzzy reasoning by genetic algorithms and its application to discrimination of myocardial heart disease. Proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference, 1756–1761 http://www.ifcc.org/ejifcc/vol14no2/140206200308n.htm

[7] J. Derisi, V. Iyer, P. Brosn, Exploring the metabolic andgenetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[8] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Bostein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[9] R. Shamir, R. Sharan, Algorithmic approaches to clustering gene expression data, in: T. Jiang, T. Smith, Y. Xu, M.Q. Zhang (Eds.), Current Topics in Computational Biology, MIT Press, Cambridge, MA, 2001

[10] G. Getz, E. Levine, E. Domany, Coupledtwo-way clustering analysis of gene microarray data, Proc. Natl. Acad. Sci. USA 97 (22) (2000) 12079–12084

[11] I.N. Sarkar, P.J. Planet, T.E. Bael, S.E. Stanley, M. Siddall, R. DeSalle, D.H. Figurski, Characteristic attributes in cancer microarrays, J. Biomed. Inf. 35 (2) (2002) 111–122.

[12] L. Wuju, X. Momiao, Tclass: tumor classi:cation system basedon gene expression pro:les, Bioinformatics 18 (2002) 325–326

[13] Y. Li, C. Campbell, M. Tipping, Bayesian automatic relevance determination algorithms for classifying gene expression data, Bioinformatics 18 (2002) 1332–1339.

[14] W. Li, I. Grosse, Gene selection criterion for discriminant microarray data analysis based on extreme value distributions, RECOMB03: Proceedings of the Seventh Annual International Conference on Computational Biology, 2003

[15] GraphViz, Graph Visualization Project, http://www.graphviz.org/

[16] M. Xiong, W. Li, J. Zhao, L. Jin, E. Boerwinkle, Feature (Gene) selection in gene expression-based tumor classi:cation, Mol. Genet. Metabolism 73 (3) (2001) 239–247

[17] X. Yao, Y. Liu, A new evolutionary system for evolving arti:cial neural networks, IEEE Trans. Neural Networks 8 (3) (1997) 694–713.

[18] J.-H. Ahn, S.-B. Cho, Combining multiple neural networks evolvedby speciation, ICONIP 2000, 2000, pp. 230 –234

[19] J.-H. Ahn, S.-B. Cho, Speciatedneural networks evolvedwith :tness sharing technique, Proceedings of the 2001 Congress on Evolutionary Computation, Vol. 1, 2001, pp. 390 –396

[20] S.-I. Lee, J.-H. Ahn, S.-B. Cho, Exploiting diversity of neural ensembles with speciated evolution, International Joint Conference on Neural Networks, Vol. 2, 2001, pp. 808–313.

[21] Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. Engl Appl Artif Intell 2013;26:2194–205

[22] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113–27

[23] Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. J Am Med Inform Assoc 2013;20:613–8.

[24] Waddell M, Page D, Shaughnessy Jr J. Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma. ACM 2005:21–8

[25] Park C, Ahn J, Kim H, Park S. Integrative gene network construction to analyze cancer recurrence using semi-supervised learning. PLoS One 2014;9:e86309

[26] Tseng C-J, Lu C-J, Chang C-C, Chen G-D. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput & Applic 2014;24: 1311–6