VOL. 13, NO. 20, OCTOBER 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

INTEGRATED FROMEWORK FOR PROGNOSIS OF CERVICAL DYSPLASIA

Shantanu S Pathak and D Rajeswara Rao
Koneru Laxmanai Education Foundation, Koneru Lakshmaiah University, Guntur, Andhra Pradesh, India
E-Mail: shantanuspathak@gmail.com

ABSTRACT

Prognosis of various diseases has been a challenge in medical field. With advances in Cognitive Sciences, Neural Networks and Machine Learning this challenge is being addressed at various levels. Cervical Dysplasia or cancer is one of the major reasons for women deaths. So, here novel framework based on reservoir computing is applied on this problem. Also, most commonly faced lack of labelled data and partial data availability in medical field is addressed by this framework. Proposed framework is compared with current machine learning models like Random Forest, Support Vector Machine, AdaBoost and multi-layer perceptron. Results over ten various parameters prove it is best in cost amongst all over this dataset. Also, it is stable in partial availability of information.

Keywords: reservoir computing, carcinoma, echo state networks, healthcare prognosis, neural networks.

1. INTRODUCTION

Prognosis of various diseases has been a challenge in medical field. With advances in Cognitive Sciences, Neural Networks and Machine Learning this challenge is being addressed at various levels. Carcinoma being one of the most sever diseases has been field of interest for research community. In literature several attempts are made to improve prognosis of carcinoma.

Here in this work, prognosis of cervical dysplasia is handled using model approach based on reservoir computing. Reservoir computing has proved to be an efficient model on solution to problems of various domains. Also, for benchmarking purpose proposed framework is compared with current models of machine learning. Models are compared based on ten various performance parameters. Additionally, models are tested for performance of datasets with partial information.

Organization of this work is classical. Literature survey follows after this section. Next, reservoir computing is discussed. Then proposed approach is presented in details. Dataset description, results and discussion are next in sequence. Finally, conclusions are presented.

2. LITERATURE SURVEY

In literature various machine learning methods for prognosis of carcinoma are discussed. Prediction of existence of carcinogenic cells is handled in [1, 2] work. This problem is relevant till days even if studied since early days. Further, Hidden Markov Model's (HMM) application in cancer prediction is discussed in work [3, 4]. These works show how classically proved HMM model is applicable even in cancer domain. Additionally, work [5] studies Random Forest approach for same problem. Random forest classifiers have proved to be efficient because of ensemble based learning. In work [6] computing techniques are applied for multi-class classification problem regarding cancer stage detection. This is also an important problem as detection of stage can help precise treatment. Feature selection problem is addressed in work [7], on other hand [8] discusses multi parametric image classification problem. Feature selection is always been important problem in all types of machine learning and decision-making problems. This work efficiently selects relevant features to improve performance of models. Another work on cancer related smoking habit to diagnosis [9]. Smoking has proved to be correlated to many types of cancers. Use of rough-set theory for incomplete information processing is demonstrated in work [10]. Similarly work [11] handles partial observations. Such processing is helpful for medical domain applications where complete information is often missing. Lack of complete information is due to various issues like privacy and time for various test reports.

In all these works, supervised learning problems are addressed assuming availability of labelled data for training. In proposed approach even, absence of such training samples can be handled. For this purpose, unsupervised approach is used to label input data. Then these labels are used to train supervised model.

3. RESERVOIR COMPUTING

It is a neural network paradigm dealing with random synaptic connections between neurons acting as reservoir. Here uniqueness is in approach of reflecting provided input amongst neurons within reservoir. Such reflections results in mapping to output pattern. This approach is popular since first proposed by authors in [12]. Figure-1 shows details of approach in graphical form.
Output of system depends on input and reservoir state.

State of the reservoir depends on history of inputs. To put it formally, say input I_t is given to receive output O_t at time 't'. Historically, I_{t-n} .. I_{t-1} was received by the reservoir. Here 'n' is the number of historical steps to be considered.

$$O_t = f(I_t) + S(n)$$

$$S(n) = \sum_{i=1}^{n} g(I_{t-i})$$

One important paradigm of reservoir computing is Recurrent Neural Networks (RNN). These are neural networks designed to handle recurring patterns. Patterns which repeat themselves over time or space can be easily detected by them. This is achieved by maintaining information regarding past patterns [13]. Even nonlinear time series can be modelled using these networks [14, 15]. Recently even pre-trained models are available to be directly be used for time series predictions [16]. In RNN output is dependent on input and hidden state maintained by the network [17]. The current hidden state is dependent on previous hidden state. So, RNNs systematically model sequential dependence of input. Assume, at time t consider y_t as output, for input x_t and current hidden state t_m. w_{ho} represents weights assigned to connection between input and hidden state, w_{hy} represent weights between hidden states h_t and h_{t-1} that represents weights for hidden state to output connection. W represents all types of weights. In simplest form these dependencies can be expressed as,

$$y_t = f(h_t, W)$$

$$h_t = g(h_{t-1}, x_t, W)$$

4. PROPOSED APPROACH

Major objective of proposed approach is to reduce dependence on need of training data. This is achieved by using unsupervised unit's output for training supervised unit. This reservoir computing based approach is novel in its application to cervical cancer problem. Also, use of reservoir computing makes enables model to handle partial information. Partial availability of information is common in medical field because of privacy and other issues.

Figure-1 shows RC based cognitive reflective multi-aspect Framework (CRMF) with all units in details. All units are Neural Network based. First unlabeled data is provided to unsupervised unit which labels it. These labels are used for training supervised reservoir computing unit. The unsupervised phase uses competitive learning approach. Here, neurons compete to match closer to the input pattern. The neuron having synaptic weights closest to input pattern wins, which decides label for given input. Initially neurons are assigned random synaptic weights.

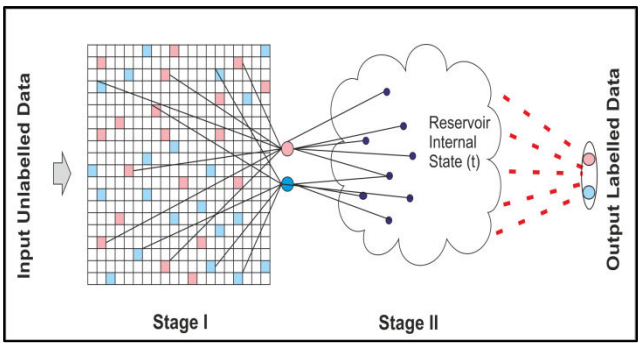

Figure 1: RC based Cognitive Reflective Multi-aspect Framework (CRMF)

In reservoir computing based unit, reservoir contains n neurons. These neurons are randomly connected to each other, ensuring sparse connections. Each connection enables reflection of input to various neurons within reservoir. Such sparse connections enable controlled reflections. Finally, there are fixed connections from reservoir to output. Output determines predicted label of input. In training phase, error between predicted and actual value is propagated for synaptic weight updating only in output connections. This is major difference between traditional neural networks and reservoir computing. In traditional neural networks error will be propagated to all layers and in reservoir computing only output layer weights are updated.

5. DATASET DESCRIPTION

5.1 Basic statistics

Here dataset Cervical of cancer from [11] is taken for experiment. There are total 858 samples, 36 features for this binary classification problem. Samples show acceptable level of class imbalance. Before experiment, data is cleaned in pre-processing stage.

Four flavours of same dataset are created for experimenting with partial information. First flavour is encoded full dataset. Here One Hot encoding applied on categorical attributes. Then two more flavours are created by removing attributes from dataset at testing phase. So, the models are tested for working on partial information. Attributes offering lowest information gain are removed one by one. So, RM (1) refers to dataset with lowest attribute removed and RM (2) means lowest two attributes removed in information gain ranking. Finally the full dataset is used as it is for comparison purpose.

Table-1. Basic statistics of dataset.

Data head	Description	Value
Cervical Cancer (Full)	Samples	858
	Features	36
Cervical Cancer (Encoded)	Samples	858
	Features	36+Encoding
Cervical Cancer (RM(1))	Samples	858
	Features	35
Cervical Cancer (RM(2))	Samples	858
	Features	34

5.2 Data pre-processing

In medical field datasets missing values are common. This problem can be efficiently handled by using mode value of an attribute. Such approach only works well if sufficient parts of values are present for given attribute. Also, in a dataset if samples from various classes differ in percentage then this add bias to classifier. It is commonly known as class imbalance problem.

In machine learning, data attributes are assumed to be completely independent of each other. Inter-dependence between attributes is termed as Co-linearity problem. It is one of the causes of poor model performance. So, removing co-linearity using Principle Component Analysis (PCA) is done here.

6. EXPERIMENT RESULTS AND DISCUSSION

Proposed framework RCIF is compared with state of the art machine learning models, AdaBoost, random forest, support vector machine (SVM) and multi layer perceptron (MLP). They are all compared based on ten standard parameters like accuracy, precision, recall, F1 Score. Table-1 shows detailed values of those parameters. Best performing figures are highlighted in bold. In Figures 2a, 2b, 2c and 2d bar chart presents visual comparison of parameters mentioned.

Analysis of results shows that MLP performs its best on RM (1) dataset. It gives its best accuracy and even best cost on it. SVM performs steadily well on all datasets, even if it's a traditional approach. This also shows that it is stable in absence of partial information.

In general, AdaBoost has proved to be a promising model. But here it fails to perform its best. Only on encoded dataset its cost is comparable with others. Random Forest model because of its randomness has proved best over multiple parameters across multiple datasets.

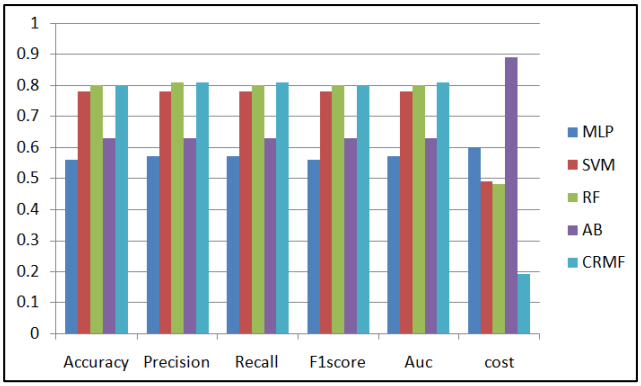

Figure-2a. Results for full dataset.

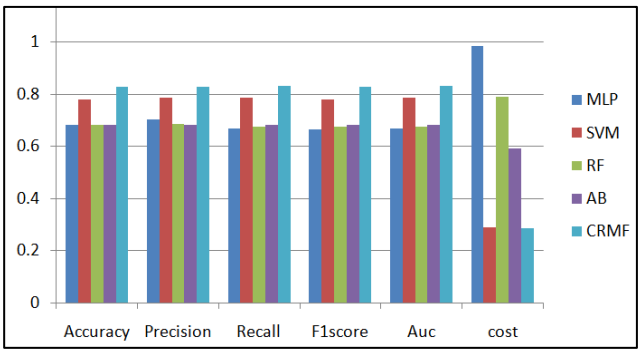

Figure-2b. Results for encoded dataset.

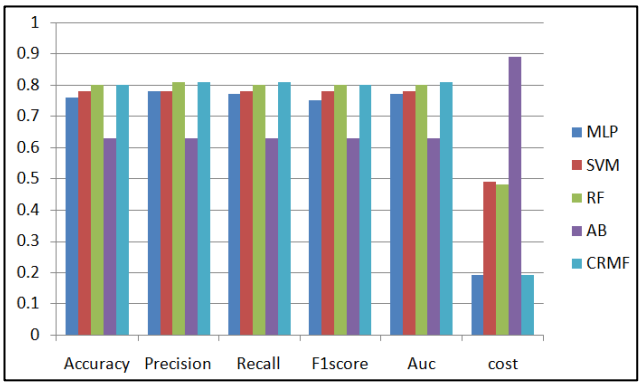

Figure-2c. Results for RM (1) dataset.

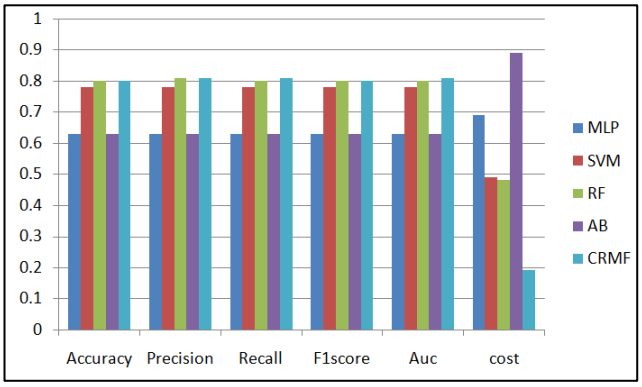

Figure-2d. Results for RM (2) dataset.

RCIF gives comparative results with minimum cost over all datasets across all parameters. It outperforms all models in terms of cost. Also, on other parameters it has best results. On few parameters best place is shared among RCIF and random forest. Overall proposed approach RCIF has upper hand over other models.

7. CONCLUSIONS

Cervical dysplasia is one of the causes of deaths in women. Prognosis of such a detrimental disease will help masses. Here novel approach based on reservoir computing is proposed to solve this problem. This approach has proved efficient and can even handle un-labelled data. Here, performance of model is judged by ten parameters against state of the art models. Additionally, all models are tested for their performance of partial information. Proposed approach has proved stable even in an environment with partial information.

REFERENCES

[1] J. A. Cruz and D. S. Wishart. 2006. Applications of machine learning in cancer prediction and prognosis. Cancer informatics. 2: 117693510600200030.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis. 2015. Machine learning applications in cancer prognosis and

prediction. Computational and structural biotechnology journal. 13: 8-17.

[3] G. Manogaran, V. Vijayakumar, R. Varatharajan, P. M. Kumar, R. Sundarasekar and C.-H. Hsu. 2017. Machine learning based big data processing framework for cancer diagnosis using hidden markov model and gm clustering. Wireless Personal Communications. pp. 1-18.

[4] S. Mukhopadhyay, I. Kurmi, S. Pratiher, S. Mukherjee, R. Barman, N. Ghosh and P. K. Panigrah. 2018. Efficacy of hidden markov model over support vector machine on multiclass classification of healthy and cancerous cervical tissues. in Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics, vol. 10501, p. 105011M, International Society for Optics and Photonics.

[5] G. Sun, S. Li, Y. Cao and F. Lang. 2017. Cervical cancer diagnosis based on random forest. International Journal of Performability Engineering. 13(4): 446-457.

[6] P. Mitra, S. Mitra and S. K. Pal. 2000. Staging of cervical cancer with soft computing. IEEE Transactions on Biomedical Engineering. 47(7): 934-940.

[7] J. Zhang and Y. Liu. 2004. Cervical cancer detection using svm based feature screening. in International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 873-880, Springer, 2004.

[8] T. Torheim, E. Malinen, K. H. Hole, K. V. Lund, U. G. Indahl, H. Lyng, K. Kvaal and C. M. Futsaether. 2017. Autodelineation of cervical cancers using multiparametric magnetic resonance imaging and machine learning. Acta Oncologica. 56(6): 806-812.

[9] T. Onega, E. Nutter, J. Sargent, J. Doherty and S. Hassanpour. 2017. Identifying patient smoking history for cessation and lung cancer screening through mining electronic health records. Cancer Epidemiology and Prevention Biomarkers. 26(3): 437-437.

[10] R. Rohmat Saedudin, H. Mahdin, S. Kasim, E. Sutoyo, I. Yanto and R. Hassan. 2018. A relative tolerance relation of rough set for incomplete information systems.

[11] K. Fernandes, J. S. Cardoso and J. Fernandes. 2017. Transfer learning with partial observability applied to cervical cancer screening. in Iberian conference on pattern recognition and image analysis. pp. 243-250, Springer, 2017.

[12] D. H. Ackley, G. E. Hinton and T. J. Sejnowski. 1985. A learning algorithm for Boltzmann machines. Cognitive science. 9(1): 147-169.

[13] J. T. Connor, R. D. Martin and L. E. Atlas. 1994. Recurrent neural networks and robust time series prediction. IEEE transactions on neural networks. 5(2): 240-254.

[14] E. Egrioglu, U. Yolcu, C. H. Aladag and E. Bas. 2015. Recurrent multiplicative neuron model artificial neural network for non-linear time series forecasting. Neural Processing Letters. 41(2): 249-258.

[15] Z. Che, S. Purushotham, K. Cho, D. Sontag and Y. Liu. 2016. Recurrent neural networks for multivariate time series with missing values. arXiv preprint arXiv:1606.01865.

[16] P. Malhotra, V. TV, L. Vig, P. Agarwal and G. Shroff. 2017. Timenet: Pretrained deep recurrent neural network for time series classification. arXiv preprint arXiv: 1706.08838.

[17] B. A. Pearlmutter. 1989. Learning state space trajectories in recurrent neural networks. Neural Computation. 1(2): 263-269.