An Optimized Learning Approach for Classifying Gene Expression Datasets

Immaculate Mercy A, Chidambaram M
PG and Research Department of Computer Science, A.V.V.M, Sri Pushpam College (Autonomous), Poondi , Thanjavur , India
PG and Research Department of Computer Science, Rajah Serfoji Govt. Arts College (Autonomous), Thanjavur, India
mercybastnj@yahoo.com
chidsuba@gmail.com

Abstract— **Machine learning algorithms have been of great utilization by various industries and have gone way beyond the way in which it has proved to yield best predictions. The work that has been taken for the analysis has found out the use of three efficient ways of processing, classifying and predicting the data. The robustness of the Enhanced Generalized sequential pattern classification (EGSP) , the Effective Random Forest Binary Tree Classification (ERFBT) and the Deep Learning classifier using a Hybrid LSTM network have all proved in terms of efficiency and effectiveness of the classification and prediction. The accuracy of the classification and the predictors seems to work well even if the size of the dataset is very large. These algorithms have proved to work well for supervised, semi-supervised and unsupervised datasets and the error rate has been reduced to the maximum when compared to all the existing algorithms. These algorithms have very clear approaches that could be very well used by the health and health informatics centres as well. The algorithm employed here has a good advantage over all the existing traditional methods that exist in the industry today.**

***Keywords*- Machine Learning, Classification, Prediction, LSTM, Random Forest, Deep Learning**

I INTRODUCTION

The current era of "Big Data" has made machine learning much easier. The deep learning algorithms have shown acceptable and achievable performances even in the trained dataset consists of millions of labeled examples. This learning concept has found its extension to unsupervised or semi-supervised learning. The reason of neural networks being accepted widely is that we currently possess computational models that are connected. Larger networks are able to achieve higher accuracy on more complex tasks. Deep learning is an approach to machine learning that has drawn heavily on our knowledge of the human brain statistics and applied math. Many machine learning problems become exceedingly difficult when the number of dimensions in the data is high which is known as cure of dimensionality, Modern deep learning provides a very powerful framework for supervised learning. By adding more layers and more units within a layer, a deep network can represent functions of increasing complexity.

Deep feed forward networks also called as feed forward neural networks or multilayer perceptrons are very essential deep learning models. The idea behind feed forward network is to approximate some function f*. In the feed forward models the information flows through the function being evaluated from *'x'* through the intermediate computations used to define f and finally the output y. Each hidden layer of the network is typically vector-valued, The dimensionality of the hidden layer determines the width of the model.

Due to its greater flexibility and acceptance by various domains in the world, deep learning has taken over which is a subset of Artificial intelligence. Artificial intelligence is the future of any domain in the world and its is rapidly

taking over. The current pandemic of COVID-19 has been predicted based on the Canadian AI model . It made use of an AI tool namely BlueDot which predicted the outbreak of infectious diseases at the end of 2019 before WHO made its announcement only by January 2020. AI can be used to track (including nowcasting) and to predict how the COVID-19 disease will spread over time and over space. The Carnegie Mellon University, used algorithms to train and predict the seasonal flu, are now being re-trained on new data from COVID-19. This clearly assures that fact the AI for sure will rule the future. It doesn't stand only on the ways for infectious disease spread it also stands at the root of prediction of any disease based on the genetic information of the humans. Even in the current case of patients who has survived and those deceased it could clearly seen that even elderly patients aged above 85 years have survived the pandemic. The reason mainly lies behind their genes and how the protein interaction happens and how the superfluous regions of the genes react. So the study of this work clearly harnesses the fact the prediction at the gene level helps us to caution and take precautionary measures to any disease that could affect in the future. The work in the future could be vast if the gene study of every national could be taken into consideration. Prediction would be easier but it requires time, effort and the policies that need to be framed by each and every government in the world.

II. RELATED WORKS

Numerous efforts have been taking place in the domains of health care, science and information technology for finding out better methodologies for working with prediction and data analytics as the data has been growing in terms of Zetta bytes each and every day. The need for careful and knowledgeable ways for exploring the unexplored data which has been accumulating over the last years has been a challenge. The machine learning algorithms seems to pave way for these types of works. Though the traditional methods cannot be promising more accurate and approximate machine learning algorithms and deep learning algorithms with advancements serves the purpose. The work that has been taken for the study is the Gene expression datasets. The motivation behind Gene expression datasets is to predict the occurrence of diseases where both supervised and unsupervised data are taken into consideration. The existing methods used for the classification and prediction has limited themselves to very few data sets , where ordinary methods like the filter and the wrapper approach has been made use of. The filter approach works fine if the accuracy factor is not considered.

The wrapper methods becomes costly in terms of computation na has a greater risk of over fitting. The direct feature selection and classification on the combined datasets does result in good accuracy. Ensemble system works fine only for limited classifiers and datasets. The Fuzzy e-means learning algorithm suffers from the selection of specific gene selection as it concentrates more on binary classification rather than sequential inputs or multiclass problems. The DFNT model combines the fisher ratio and neighborhood rough set for dimensionality reduction for better classification accuracy. The Random Forest algorithm used a better classification for the Lung Cancer, Colon and Prostate Tumor datasets.

Likewise the microarray datasets are effective only for limited datasets . The multi-objective heuristic algorithm brings out the classification based on two attributed namely Accuracy being maximized and the number of features being minimized. The gene selection algorithm which made use of the group lasso classification suffers from detecting complex biological gene pathways. Also if the group size is not uniform it does not contribute to the factor of accuracy. The Microarray datasets achieved higher accuracy with a local optimum. A novel RF classification approach dealt with high dimensional data using subspace feature sampling method and feature value searching. However this method was only concerned with lowering the prediction error but has not unraveled the randomness and diversity of the forest. The Gini and permutation measures for ranking the candidate predictors had a clear cut-off for distinguishing important and non-important variables. But it has suffered from the fact that the computation time for larger datasets suffered from the accuracy. The cuckoo search algorithm classifies the binary datasets using the ELM learning algorithm which was used to train the single layer feed forward neural networks in classification field.

The prediction of breast cancer used a deep convolutional network which had a better computation time. Though the efficiency was proved by many of the existing methods the amount of the trained and untrained datasets were all

limited. So the scope as well limited. To overcome these limitations and to overcome the curse of dimensionality a new framework has been proposed which uses various algorithms in a hybrid way.

III METHODOLOGY

The motivation for the Optimized learning approach for classifying Gene expression datasets is three fold. First, the need for achieving a better classification on extremely large datasets. Two, the need for achieving better computation and accuracy. Three, an absolute prediction over the classifiers. This three- fold approach has created a need for classifying high dimensional data. The data taken for the analysis are primary gene data sets, the chromosomal gene expression datasets and the GSM datasets. Each of these sets of data has their own characteristics and each of these datasets needs to be cleaned in a different way. Likewise, the feature selection methods and the classification methods are different as they differ in their primary structure.

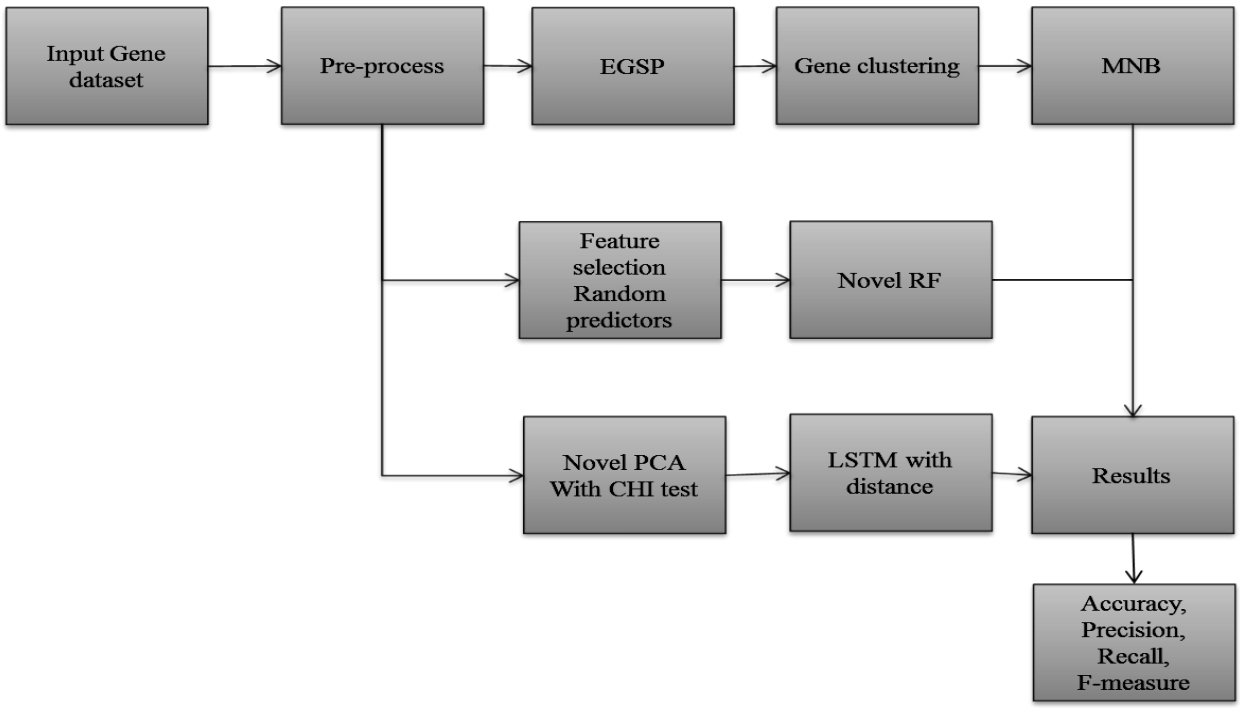

Figure 1 Overall Flow of the Process

Three approaches has been used as has been discussed on three different datasets which are all gene datasets and for the species Homo sapiens.

A. EGSP

The EGSP is the Enhanced Generalized Sequential pattern classification. The motivation for this approach is twofold. This two- fold approach has created a channel for the classification in a niche way. These sequence datasets are acquired from the publicly available gene databases. This approach makes use of the Modified Naïve Bayes algorithm. This algorithm takes the clustered sequences as input, where the training and the test sets are initialized. A count for each class is computed; with these computations the probabilistic components are computed. The motivation behind this work is that in the earlier algorithms there were certain challenges where

appropriate classification was not achieved on unlabelled data sets. The working of the EGSP and the Modified Naïve Bayes algorithm has proved to be beneficial since all the crucial factors for the classification were considered. The choosing of the threshold value is a new inclusion to the already existing algorithms where the traditional algorithms have failed to perform. As the datasets are surplus, the deficiencies that have been encountered in the earlier methodologies are well harnessed.

The EGSP brings out a prediction model for working with the sequence datasets. It proceeds by the way of generalizing the datasets which are supervised and unsupervised in nature. The sequences are generated based on the threshold value which is then followed by applying the EGSP algorithm which brings out the sequential pattern from the pruned sequences. The extracted sequential patterns from the pruned sequences are then clustered using the clustering algorithm. The MNBC algorithm (Modified Naïve Bayes) computed the probabilistic components for each class. The accuracy obtained is well beyond the traditional classification algorithms. Using this algorithm has drastically reduced the computational costs. The statistical measures have undoubtedly proved that the proposed algorithm is way ahead of all the existing methodologies.

B. ERFBT

The ERFBT (Effective Random Forest Boot strap Technique) aims at predicting the abnormalities in human chromose-17 by means of appropriate wrapper and filter methods. Initially the pre-processing is done for achieving pure data. The feature selection process is carried out with the random forest classifier until the condition satisfies for growing the tree. The data is split up for best prediction. The bootstrap samples and randomized trees learn from the sample. It is clear from the experimental results that the ERFBT outperforms the other algorithms. The abnormalities in human chromosome -17 are achieved by the effective use of the technique. The accuracy that has been obtained in this classification is far better than the existing methods. The computation time is also very less.

C. Hybrid LSTM Network

A deep learning system is incorporated for the classification of differentially expressed genes which houses hybrid methods for the classification process. The Long Short-term Memory (LSTM) is employed with k-nearest Neighbor algorithm to achieve classification to its precision. The classification further leads to enhanced prediction method. This work is well supported by the use of feature selection by the use of Hybrid Principal Component analysis and the Chi square test. The features that are obtained are ranked by these scores and the datasets which has the highest scores are further taken for training. The accuracy of the classification and the prediction that has been achieved surpasses all the existing methods.

IV. IMPLICATIONS

The main motivation behind working with these gene sequences is that prediction for occurrences of diseases if known well on hand appropriate measures could be taken to prevent and suppress the severity of the diseases, for which the human genes plays an important role. The human gene has been engineered in such a way that everything has been coded and that coding needs to be analyzed. Moreover there are various kinds of gene databases which are publicly available and they are frequently uploaded. Most of these datasets are structured, some unstructured, some labeled, may unlabelled, some supervised many semi-supervised or unsupervised. So all these range of datasets need to considered for the study. The proposed works implies that the classification and the prediction are harnessed to the maximum level of accuracy. So these models could be very well considered for the prediction in the future.

V. CONCLUSION

All the above methods have clearly supported the fact that has beneficial approaches for classification for the high dimensional structured and unstructured data. As the need for classification and prediction is on the rise the proposed methodologies aims at bringing out a prediction model that could serve the analysis in the domain of biomedical research. These machine learning and deep learning trends seems to have a clear edge in analyzing biomedical data due to the rising complexities. Predominantly the gene expression methods are in need of better predictors that could serve the health industry. So by employing the discussed and the discussed methods this starvation could be dealt with.

REFERENCES

[1]. Heba Abusamra. "*A comparative study of featue selection and clasification methods for gene expression data of glioma*", Procedia Science Direct , Elsevier Issue.10.1016/j.procs.2013.10.003. For Conference.

[2]. Jia Lv,Qinke Peng,Xiao Chen, Zhi Sun , "*A multi-objective heuristic algorithm for gene expression microaray data classification*", Elsevier, Expert Systems with Applications 59(2016)13-19.

[3]. Krisztian Buza, "*Classification of Gene Expression dat: A Hubness-aware semi-supervised approach*", Elsevier, Computer Methods and Programs in Biomedicine 127(2016) 105-113.

[4]. Hung-Yi Lin, "*Gene Discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification",*Elsevier, Applied Soft Computing 48(2016) 683-690..

[5]. Sara Tarek,Reda Abd Elwahab,Mahmoud Shoman, "*Gene Expression based cancer classification*", Egyptian Informatics Journal 2016.

[6]. Devi Arockia Vanitha ,Devaraj D,Venkatesulu, "*Gene Expression Data classification using support Vector Machine and Mutual Information–based Gene selection*", Procedia Computer science 47(2015)13-21.

[7]. Konstantina Kourou Costas Papaloukas,Dimitrois I.Fotiadis, "*Intergration of pathway Knowledge and Dynamic Bayesian Networks for the prediction of Oral Cancer Recurrence*", IEEE 2016.

[8]. Wang, Y., X. Li, and R. Ruiz, *Weighted general group lasso for gene selection in cancer classification.* IEEE transactions on cybernetics, 2018. **49**(8): p. 2860-2873.

[9]. Li, J., W. Dong, and D. Meng, *Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information.* IEEE/ACM transactions on computational biology and bioinformatics, 2017.

[10]. Yang, Y. and H. Zou, *A fast unified algorithm for solving group-lasso penalize learning problems.* Statistics and Computing, 2015. **25**(6): p. 1129-1141.

[11]. Liu, C. and H. San Wong, *Structured Penalized Logistic Regression for Gene Selection in Gene Expression Data Analysis.* IEEE/ACM transactions on computational biology and bioinformatics, 2017.

[12]. Wu, H.-C., X.-G. Wei, and S.-C. Chan, *Novel Consensus Gene Selection Criteria for Distributed GPU Partial Least Squares-based Gene Microarray Analysis in Diffused Large B cell Lymphoma (DLBCL) and related findings.* IEEE/ACM transactions on computational biology and bioinformatics, 2017.

[13]. Xu, J., et al., *A Novel Deep Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data.* IEEE Access, 2019. **7**: p. 22086-22095.

[14]. Chen, K.-S., et al., *Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome.* Nature genetics, 1997. **17**(2): p. 154.

[15]. Nguyen, T., et al., *Medical data classification using interval type-2 fuzzy logic system and wavelets.* Applied Soft Computing, 2015. **30**: p. 812-822.

[16]. Sahmadi, B., et al. *A modified firefly algorithm with support vector machine for medical data classification*. in *Computational Intelligence and Its Applications: 6th IFIP TC 5 International Conference, CIIA 2018, Oran, Algeria, May 8-10, 2018, Proceedings 6*. 2018. Springer.

[17]. Shankai Yan and Ka-Chun Wong, "GESgnExt: Gene Expression Signature Extraction and Meta-analysis on Gene Expression Omnibus" in Journal of Biomedical and Health Informatics vol. 24, issue 1, January 2020

[18]. Su-Ping Deng,Liu Zhu and De-Shuang Huang ,"Predicting Hub Genes Associated with Cervical Cancer through Gene Co-Expression Networks " *IEEE Transactions on Computational Biology and Bioinformatics*, Vol 13, No.1, January 2016.

[19]. Lei Chen, Xiao Yong Pan, Yu-Hang Zhang, Min Liu, Tao Huang, Yu-Dong Cai , "Classification of widely and rarely expressed genes with recurrent neural network" *Computational and Structural Biotechnology Journal* December 2018.

[20]. Xing-Lin Hsu, Po-Yu Huang and Dung-Tsa Chen , "Sparse PCA for Cancer Calssification " in *PMC – US National Library of Medicine, National Institute of Health*, Vol 3, June 2014.

[21]. P.Paokanta, "β- Thalassemia Knowledge Elicitation Using Data Engineering: PCA , Pearson"s Chi square and Machine Learning " *International Journal of Computer Theory and Engineering* Vol 4, No.5, October 2012.

[22]. D. S. Huang and C. H. Zheng, "Independent component analysis based penalized discriminate method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15,

[23]. World Health Organization, *World Cancer Report 201*4, pp, Chapter 5.12, 2014.