PREDICTION OF GENETIC DISEASES BASED ON DNA

M. Vineela , Associate Professor, Department of CSE, vineela_m_99@yahoo.com

G.Niharika, B.Tech, Department of CSE, niharikagangadhar5@gmail.com

B. Anitha, B.Tech, Department of CSE, anithabari328@gmail.com

P. Harika, B.Tech, Department of CSE, harikap2021@gmail.com

M.Priyanka, B.Tech, Department of CSE, priyankamalkapeta505@gmail.com

ABSTRACT: The invention of new technologies in the filed of genetic diseases resulted the easiness to treat the genetic diseases. Among all the foremost daunting tasks within the post-genomic period , one among those is the detection of genes that cause diseases from an outsized amount of genetic data. Complex diseases often present a highly heterogeneous genotype, which makes it difficult to acknowledge biological markers. Machine learning algorithms are commonly want to define such markers, but their success depends heavily on the dimensions and quality of the info available. The machine learning area, which mainly aims to create algorithms that improve with practice, promises to permit computers to assist people, analyse big and complex data sets. we developed a supervised methodology of machine learning to predict complex genes that cause diseases and the designed algorithm was experimented that Gene Ontology (GO)-trained machine learning classifiers can enhance and identify the genes that are involved in complex diseases. The analyzer for genetic diseases, Gentic Diseases Analyzer (GDA) using machine learning have been formulated by using the hybrid model of PCA, Regression, Random Forest, Decision tress algorithms. The GDA was experimented with PCA and Random Forest algorithms and the results are compared. The P-GDA model is provided the accuracy as 97.34% and sensitivity as 96.45% for the GEO dataset. The findings of machine learning approaches and their practical implementation was discussed for the study of genetic and genomic data sets.

1. INTRODUCTION

The last 10 years have seen remarkable development in the discovery of genes associated with types of neurological disorders associated with Medellin. Through human gene discovery is important for the affected patients, for the researchers focused on the disease and for the broader neuroscience community [10]. These gene discoveries may however be even more significant when collectively considered. First and foremost, the development of rare disease genetics has shown that central and peripheral nervous system dysfunction is a common result of genetic disorders with approximately 50% of all rare diseases (3000-3500 conditions) having some type of neurological abnormality [3]. The fact that neurological disorders display extraordinary genetic variation has also become apparent. It has become clear, eventually, that variable expressiveness and a typical expressions are not the exception, but the rule when contemplating neurogenetic disorders [4]. Such

findings have led us to the idea that genes associated with neurogenetic conditions needed something "extra" to exist. The machine learning field cares with developing and applying computer algorithms which improve with experience [6]. Machine learning algorithms use input file to perform as well as learning the way to recognize gene patterns in DNA.

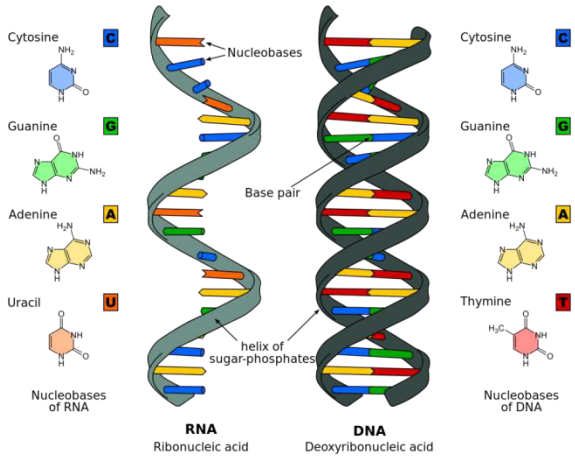

Fig.1: Example figure

Many risks are associate while analyzing with the machine learning tools for genetic diseases analysis [10]. The study address whether the HSQ-23 successfully identify the people who will enjoy PHC consultations. Data from the baseline survey (n = 2056) has been used to analyze the measurement and the properties of the SQ-33 in order to determine the scaling properties for the complete range of subjects.

2. LITERATURE REVIEW

2.1 Learning Deep Gradient Descent Optimization for Image Deconvolution

As an integral component of blind image deblurring, non-blind deconvolution removes image blur with a given blur kernel, which is essential but difficult due to the ill-posed nature of the inverse problem. The predominant approach is based on optimization subject to regularization functions that are either manually designed, or learned from examples. Existing learning based methods have shown superior restoration quality but are not practical enough due to their restricted and static model design. They solely focus on learning a prior and require to know the noise level for deconvolution. We address the gap between the optimization-based and learning-based approaches by learning a universal gradient descent optimizer. We propose a Recurrent Gradient Descent Network (RGDN) by systematically incorporating deep neural networks into a fully parameterized gradient descent scheme. A hyper-parameter-free update unit shared across steps is used to generate updates from the current estimates, based on a convolutional neural network. By training on diverse examples, the Recurrent Gradient Descent Network learns an implicit image prior and a universal update rule through recursive supervision. The learned optimizer can be repeatedly used to improve the quality of diverse degenerated observations. The proposed method possesses strong interpretability and high generalization. Extensive experiments on synthetic benchmarks and challenging real-world images demonstrate that the proposed deep optimization method is effective and robust to produce favorable results as well as practical for real-world image deblurring applications.

2.2 An incremental learning approach for restricted Boltzmann machines

Determination of model complexity is a challenging issue to solve computer vision problems using restricted boltz-mann machines (RBMs). Many algorithms for feature learning depend on cross-validation or empirical methods to optimize the number of features. In this work, we propose an learning algorithm to find the optimal model complexity for the RBMs by incrementing the hidden layer. The proposed algorithm is composed of two

processes: 1) determining incrementation necessity of neurons and 2) computing the number of additional features for the increment. Specifically, the proposed algorithm uses a normalized reconstruction error in order to determine incrementation necessity and prevent unnecessary increment for the number of features during training. Our experimental results demonstrated that the proposed algorithm converges to the optimal number of features in a single layer RBMs. In the classification results, our model could outperform the non-incremental RBM.

2.3 Fuzzy soft set based classification for gene expression data

Classification is one of the major issues in Data Mining Research fields. The classification problems in medical area often classify medical dataset based on the result of medical diagnosis or description of medical treatment by the medical practitioner. This research work discusses the classification process of Gene Expression data for three different cancers which are breast cancer, lung cancer and leukemia cancer with two classes which are cancerous stage and non cancerous stage. We have applied a fuzzy soft set similarity based classifier to enhance the accuracy to predict the stages among cancer genes and the informative genes are selected by using Entopy filtering.

2.4 Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients

The liver is considered an essential organ in the human body. Liver disorders have risen globally at an unprecedented pace due to unhealthy lifestyles and excessive alcohol consumption. Chronic liver disease is one of the principal causes of death affecting large portions of the global population. An accumulation of liver-damaging factors deteriorates this condition. Obesity, an undiagnosed hepatitis infection, alcohol abuse, coughing or vomiting blood, kidney or hepatic failure, jaundice, liver encephalopathy, and many more disorders are responsible for it. Thus, immediate intervention is needed to diagnose the ailment before it is too late. Therefore, this work aims to evaluate several machine learning algorithm outputs, namely logistic regression, random forest, XGBoost, support vector machine (SVM), AdaBoost, K-NN, and decision tree for predicting and diagnosing chronic liver disease. The classification algorithms are evaluated based on various measurement criteria, such as accuracy, precision, recall, F1 score, an area under the curve (AUC), and specificity. Among the algorithms, the random forest algorithm showed better performance in liver disease prediction with an accuracy of 83.70%. Furthermore, the random forest algorithm also showed better precision, F1, recall, and AUC metrics. Hence, random forest is considered the best algorithm for early liver disease prediction.

2.5 Deep Network-Based Feature Selection for Imaging Genetics: Application to Identifying Biomarkers for Parkinson's Disease

Imaging genetics is a methodology for discovering associations between imaging and genetic variables. Many studies adopted sparse models such as sparse canonical correlation analysis (SCCA) for imaging genetics. These methods are limited to modeling the linear imaging genetics relationship and cannot capture the non-linear high-level relationship between the explored variables. Deep learning approaches are under explored in imaging genetics, compared to their great successes in many other biomedical domains such as image segmentation and disease classification. In this work, we proposed a deep learning model to select genetic features that can explain the imaging features well. Our empirical

study on simulated and real datasets demonstrated that our method outperformed the widely used SCCA method and was able to select important genetic features in a robust fashion. These promising results indicate our deep learning model has the potential to reveal new biomarkers to improve mechanistic understanding of the studied brain disorders.

3. IMPLEMENTATION

In existing work we used the quantile method for normalization of background correction with "normexp". They have employed SVM, LDA algorithms with the Hierarchical based Euclidian distance method..

Drawbacks

- They have limited the testing space by selecting two genes for classification

The aim of this research is typically to identify the genetic variations that are occurring in DNA which may directly or indirectly lead to the increased risk of diseases. The proposed GDA is designed as Hybrid method using PCA, Regression, Random Forest, Decision tress algorithms

Advantages

- No limited for classification

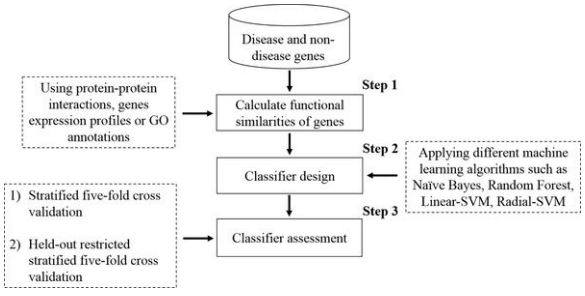

Fig.2: System architecture

4. ALGORITHMS

Here in this paper we are used algorithms like PCA, regression, Random Forest algorithms and decision tree.

PCA:

PCA is an unsupervised machine learning algorithm that attempts to reduce the dimensionality (number of features) within a dataset while still retaining as much information as possible. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover, PCA is an unsupervised statistical technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.

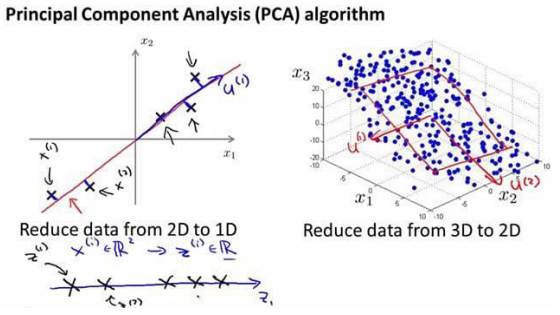

Fig.3: PCA model

Logistic regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a

given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X. Contrary to popular belief, logistic regression IS a regression model.

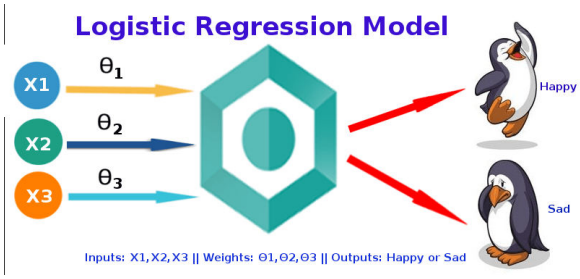

Fig.4: LR model

Random forest:

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random Forest is a supervised machine learning algorithm made up of decision trees. Random Forest is used for both classification and regression—for example, classifying whether an email is “spam” or “not spam”.

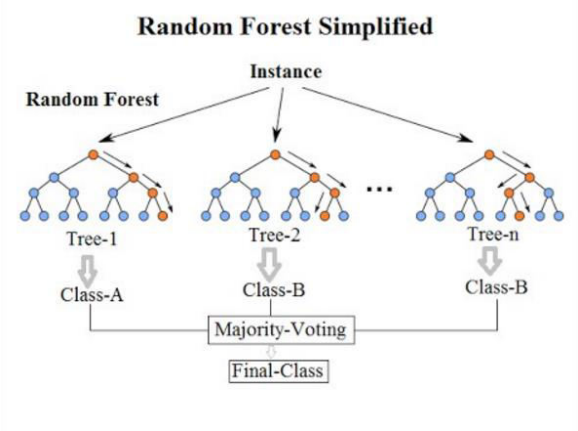

Fig.5: RF model

Decision tree:

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

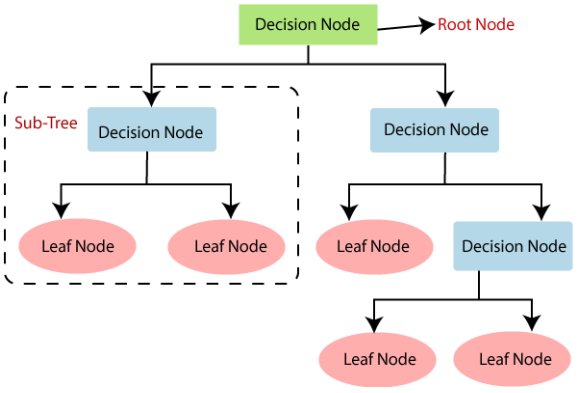

Fig.6: Decision tree model

5. EXPERIMENTAL RESULTS

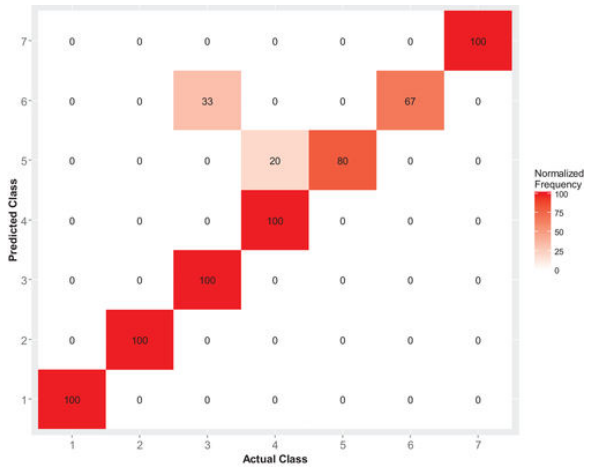

Fig.7: Output screen

	Class						
	1	2	3	4	5	6	7
Sensitivity	1	1	1	1	0.8	0.67	1
Specificity	1	1	0.91	0.95	1	1	1
Balanced accuracy	1	0.96	0.98	0.9	0.83	1	

Fig.8: Output screen

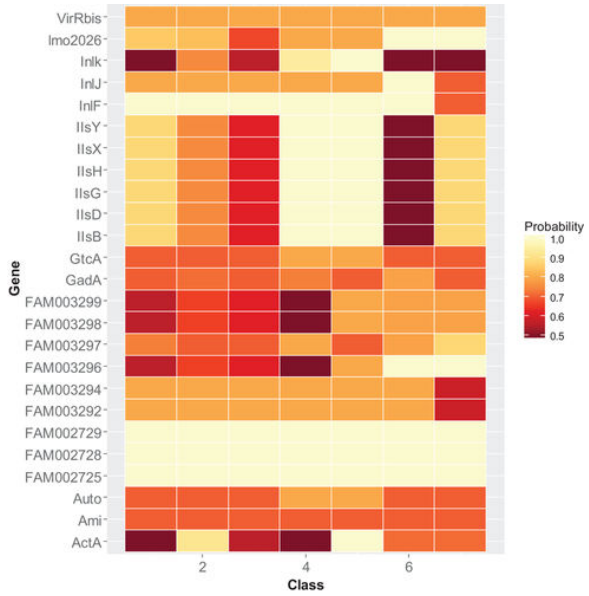

Fig.9: Output screen

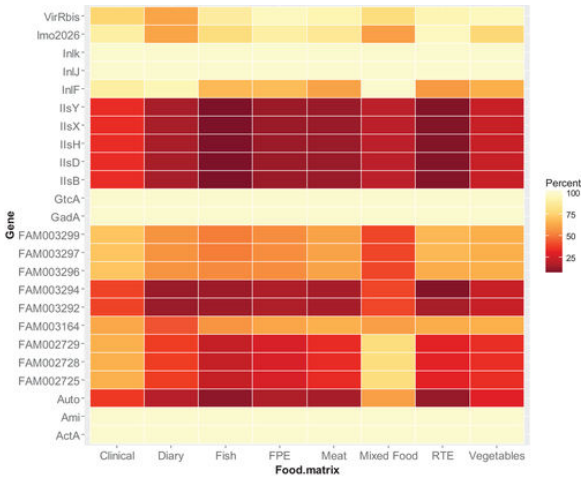

Fig.10: Output screen

6. CONCLUSION

The machine learning for genetic diseases have been analyzed. The hybrid model for analyzing genetic diseases has been proposed. The proposed GDA model has been designed using hybrid approach of PCA, Regression, Random Forest, Decision tress algorithms. The GDA was experimented with PCA and Random Forest algorithms for comparative analysis. The P-GDA model is provided the accuracy as 97.34% and sensitivity as 96.45% for the GEO dataset. The accuracy of P-GDA is higher as 3.9% and 6.17% than PCA and Random Forest algorithms respectively. The sensitivity is also outperformed as 2.2% and 2.8% than PCA and Random Forest algorithms respectively.

7. FUTURE WORK

The Future Work shall focus on the development of multi dimensional algorithm for prediction of genetic diseases.

REFERENCES

1. Alshamlan, H.M., Badr, G.H. and Alohali, Y.A., 2015. Genetic Bee Colony (GBC) algorithm, Computational Biology and Chenistry, 56, pp.49-60.

2. Hameed, 2017. Use of a combination of statistical filters and a GBPSO-SVM algorithm. PloS, 12(11), p.e0187371.

3. D. Gong, Z. Zhang, Q. Shi, A. van den Hengel, C. Shen and Y. Zhang, "Learning Deep Gradient Descent Optimization for Image Deconvolution," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2020.2968289.

4. Jongmin Yu, Jeonghwan Gwak, Sejeong Lee and Moongu Jeon, "An incremental learning approach for restricted boltzmann machines," 2015 International Conference on Control, Automation and Information Sciences (ICCAIS), Changshu, 2015, pp. 113-117, doi: 10.1109/ICCAIS.2015.7338643.

5. Kalaiselvi, N. and Inbarani, H.H., 2013. Fuzzy soft set based classification for gene expression data. arXiv preprint arXiv:1301.1502.

6. M. A. Kuzhippallil, C. Joseph and K. A, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 778-782, doi: 10.1109/ICACCS48705.2020.9074368.

7. M. Kim, J. H. Won, J. Hong, J. Kwon, H. Park and L. Shen, "Deep Network-Based Feature Selection for Imaging Genetics: Application to Identifying Biomarkers for Parkinson's Disease," 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020, pp. 1920-1923, doi: 10.1109/ISBI45749.2020.9098471.

8. Oh, D.H., Kim, I.B., Kim, S.H. and Ahn, D.H., 2017. Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning. Clinical Psychopharmacology and Neuroscience, 15(1), p.47.

9. Pandiaraj, S., Sudalai Muthu, T., Prioritization of replica for replica replacement in data grid, International Journal of Recent Technology and Engineering, 2019, Vol: 7, Issue: 5, pp. 245-248.

10. Ranjana, P., Lakshmi Sridevi, S., Sudalai Muthu, T., Vikram Gnanaraj, V., Machine Learning Algorithm in Two wheelers fuel Prediction, Proceedings of 1st International Conference on Innovations in Information and Communication Technology, ICIICT 2019, 2019.

11. Rohini, A., Sudalai Muthu, T., A weight based scheme for improving the accuracy of relationship in social network, International Journal of Innovative Technology and Exploring Engineering, 2019, Vol: 8, Issue: 11, pp. 3040-3043.

12. Rameshkumar, K., Sambath, M., Ravi, S.,2013. Relevant association rule mining from medical dataset using new irrelevant rule elimination technique, 2013 International Conference on Information Communication and Embedded Systems, ICICES 2013.