Prediction Of Lipopolysaccharides Simulation Responsiveness On Gene Expression Profiles Of Major Depression Disorder Affected Cases Using Machine Learning

Karthik Sekaran, M. Sudha

Abstract: Major Depressive Disorder is an acute-form of mental illness. It interferes in the personal life, education, eating and sleeping habits of a person affected by depression. The factors that cause depression are generally identified as environmental, genetic and other psychological reasons. Medications like anti-psychotic drug treatment and counseling are said to be responsive in controlling the mental condition for a short-term. But the current treatment methods are not effective for the patients, living with prolonged depression periods. Gene therapy gets its momentum on medical diagnostic procedures to treat the patients with handful strategies. Lipopolysaccharides, a kind of endotoxins presents in the outer membrane of gram-negative bacteria could cause potential threats to human body. In this work, the responsiveness of lipopolysaccharides simulated in blood of patients with depression over normal people is analyzed through their gene expressions. The samples are collected from Gene Expression Omnibus repository. A hybrid feature selection technique is proposed to select the biomarker genes of depression. Experimental results revealed the significant genes affected to Lipopolysaccharides simulation that discriminates the samples accurately. Machine Learning algorithms are employed to train and classify the data. This system finds 100% accurate classification of the normal and depression samples with the identified gene biomarkers.

Index Terms: Cat Boost Algorithm, Gene Expression, Machine Learning, Major Depressive Disorder, Molecular Diagnostics

1. INTRODUCTION

MAJOR Depressive Disorder (MDD) (or) Depression is a mental illness, with around 300 million people affected worldwide. Every year, more than 8, 00,000 deaths occur due to depression. It is the second leading cause of death in the age group of people lies between 15-29 years. Even though the symptoms of depression is clear, the patients receiving treatment is often less due to lack of medical facilities and trained healthcare professionals [1]. Common factors such as genetic, psychosocial and stress are the major implications of depression [2]. After the successful completion of Human Genome Project (HGP) in 2003, many gene sequencing technologies are developed [3]. High throughput sequencing technologies made the task easier and ends up in the generation of massive amount of genomic data [4]. In molecular diagnosis, gene biomarker identification is a crucial process. But, due to the complex representation of genetic data, manual evaluation becomes impossible [5]. Computational methodologies are proposed and developed in the view of identifying the biomarkers of complex. It reveals the significant genes that could cause adverse effects on the health condition. Molecular diagnostic systems extracts useful pattern from the patient's medical data for accurate diagnosis [6][7]. Feature selection is a decisive process. It directly related with the stability of a learning model. Many techniques are proposed to select best features from the data. Filter, Wrapper and Embedded method covers up most of the algorithms in this regard [8]. Many gene biomarker identification frameworks are developed for various diseases. Generally, cancer related microarray gene expression data analysis is performed in most of the cases by the contributors [9] [10] [11]. Scientific studies promises that psychiatric disorders has strong genetic link associated with its cause and influence in a person [12] [13]. But very limited theories support this fact as it has not been widely adopted and confined within a bound. Therefore, in order to highlight the need and importance of a systematic tool for molecular diagnostics, a computational model is proposed to identify the risky biomarkers of major depressive disorder from the gene expression profiles of the patients. A hybrid boosting-wrapper technique is fused to select optimal gene subset. Machine learning algorithms are employed to train and classify the samples effectively.

- *Karthik Sekaran is currently pursuing research program (Ph.D) in Computational Bioinformatics in Vellore Institute of Technology, India, PH-9597195110. E-mail: skarthik@vit.ac.in*
- *M. Sudha, Associate Professor is currently working in Vellore Institute of Technology, India,. E-mail: msudha@vit.ac.in*

2 MATERIALS AND METHODS

2.1 Dataset Information

This experiment is conducted with the dataset collected from Gene Expression Omnibus (GEO) repository, managed by National Center for Bioinformatics Institute (NCBI). The accession number of the dataset is GSE19738 [14]. The details about the samples are given as follows, 34 samples of normal group of peoples, 33 samples of MDD patients without LPS treatment and 33 samples of MDD patients after LPS treatment. The dataset is separated into two with different classes. In dataset I, the samples of normal people and MDD patients without LPS treatment is compiled and in dataset II, the samples of normal people and MDD patients with LPS treatment is fused together.

2.2 Data Preprocessing

The dataset contains 41000 gene probes as features out of which, 27,097 genes are "null" valued. In order to maintain the stability of the data, all the features with null values are eliminated. The processed dataset retains 13903 features.

2.3 Gene Biomarker Selection

The gene selection technique follows a hybrid of boosting and wrapper technique. CatBoost Algorithm is a boosting model applied on the features to calculate the score of each variable through feature learning method [15]. Top 100 ranked gene features are selected as informative. Recursive Feature Elimination (RFE) ranks the informative genes selected from CatBoost and produces optimal gene subset [16]. The number of gene probes is reduced from 13903 to 26 of dataset I and 13903 to 1 of dataset II after applying CatBoost-RFE. The top-ranked genes are the biomarkers that discriminates the samples with better accuracy. In Table 1 and 2, the score and rank of each feature selected by CatBoost-RFE is given. The rank is calculated based on the score and higher the score the rank is better.

Algorithm 1: CatBoost – RFE Algorithm

Input: $\{(Xi, Yi)\}_{i=1}^{n}, K$
σ = random perm [1, m]
N_j in 0 for j = 1...m;
for p in 1 to K do
    for j in 1 to m do
        q_j <- y_j - $N_{\sigma(j)-1}$ (X_j);
    for j in 1 to m do
        λN = Modell((x_i,q_i), σ(l) <- j);
        N_j = N_j + λN;
return N_m
Input : Training Data X
    Set of n attributes Y = { a1, a2,..., an}
    A (X, Y)
Output: Test Data Y
    Rank of each feature R
for i in 1 to n do
    Rank Y with A(X,Y)
    f ∗ ← final ranked feature in Y
    R(n − j + 1) ← f ∗
    Y ← Y − f ∗
End

TABLE 1
Gene scores of Dataset-I

Score	Gene Probe ID	Score	Gene Probe ID
0.279	A_23_P154070	0.214	A_23_P47709
0.273	A_32_P87697	0.214	A_23_P255869
0.273	A_23_P203351	0.214	A_23_P371613
0.257	A_23_P214678	0.198	A_24_P246787
0.253	A_32_P207243	0.195	A_23_P11705
0.24	A_23_P353742	0.189	A_24_P371670
0.233	A_23_P156531	0.188	A_24_P917457
0.233	A_24_P237686	0.186	A_23_P212383
0.224	A_23_P93543	0.186	A_24_P14485
0.22	A_24_P58727	0.18	A_32_P6562
0.214	A_23_P166196	0.176	A_24_P41530
0.214	A_23_P133814	0.176	A_24_P455972
0.214	A_24_P396375	0.172	A_24_P932388

TABLE 2
Gene scores of Dataset-II

Score	Gene Probe ID
1	A_23_P109436

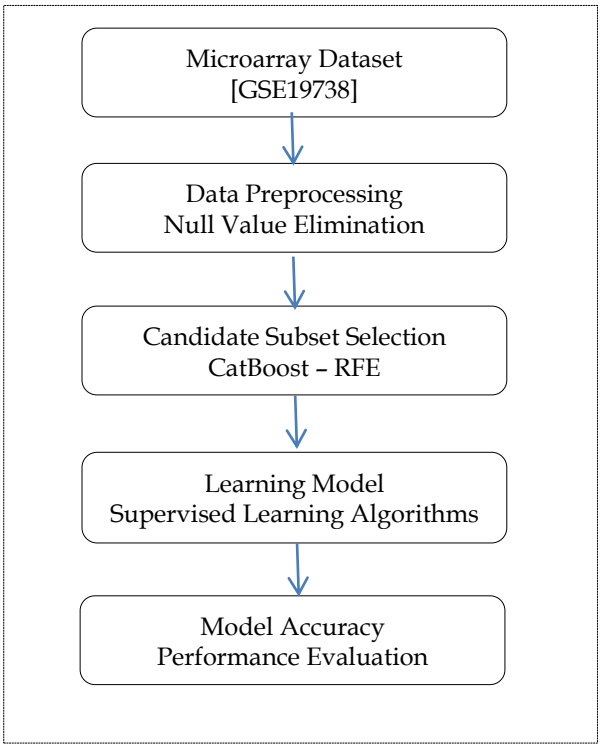

Fig 1: *General Workflow of the proposed model*

2.4 Classification

Machine Learning is a subfield of Artificial Intelligence. It empowers the process of revealing useful insights and hidden patterns from the data with its intelligent learning algorithms [17][18]. In this system, to evaluate the gene subset selected by the proposed algorithm, few learning models are employed. Bayes Net (BN), Support Vector Machines (SVM), Random Forest (RF), Back Propagation Neural Network (BPNN) and Linear Discriminant Analysis (LDA) are the algorithms discriminates the samples from the given feature subset through pattern identification.

3 RESULTS AND DISCUSSION

A gene selection technique is proposed in this work to identify the potential biomarker of MDD from gene expression profiles. CatBoost-RFE model revealed the best gene feature subset and is evaluated with learning models through k-Fold cross validation (k=10). The performance of the algorithms are validated with four model evaluation metrics namely accuracy, true positive rate (TPR), false positive rate (FPR) and f-score. The results are projected in Table 3 of dataset I. The metrics calculated for dataset II shows 100% result in all benchmarked algorithms.

TABLE 3
Performance of the proposed method for dataset I

Algorithms	Acc (%)	TPR (%)	FPR (%)	F-Score (%)
Bayes Net	95.5	95.5	4.6	95.8
RF	91.0	91.0	9.1	91.0
SVM	73.1	73.1	27.2	72.7
BPNN	71.6	71.6	28.7	71.3
LDA	70.1	70.1	29.8	70.1

The significance of the proposed work is highlighted from the results obtained. For dataset I, 26 gene biomarkers are identified and dataset II, 1 gene is selected out from 13903 genes. The experimental result shows that for dataset II, the gene A_23_P109436 identified by CatBoost-RFE accurately classifies the dataset with 100% precision. This biomarker is directly affected by the LPS treatment provided to MDD patients that discriminates the normal and depressed peoples. Machine Learning empowers wide range of applications such as weather forecasting [19] [20] [21], drug suggestion system [22], disease diagnosis from clinical trials, electronic health records [23] [24] [25] [26], sports success prediction [27] etc. More effective methods improve the predictability of the learning models. Deep learning models performs intense computation on critical applications such as image reconstruction, medical image processing and satellite image analysis. These models provide better opportunity to explore more information from complex genetic structures of heterogeneous species for effective disease diagnosis.

4 CONCLUSION

In this paper, an effective hybrid feature selection technique is proposed to select the predictor gene subset. The identified biomarkers were evaluated with supervised learning algorithms. CatBoost-RFE hybrid feature selection model identified best gene features from both the datasets. Moreover, the single gene identified in dataset II accurately discriminates the samples of both classes. This gene is responsive over Lipopolysaccharides simulation treatment on patients affected by major depressive disorder. On comparing with dataset I, the samples of MDD patients without LPS simulation lags in its results against the gene identified in dataset II. The outcome of this experiment clearly states the gene “A_23_P109436” is identified as the potential biomarker of MDD. This marker could act as the prognostic evidence to treat the patients with depression. The understanding of human genetic pattern could significantly improve the diagnostic procedures of complex diseases. In future, super-intelligent algorithms like deep learning models empower the development of computational diagnosis systems. More advancement in biomedical related fields such as pharmacogenomics, gene pathway analysis and structural drug discovery combined with computational intelligence will paves a new pathway towards “Precision Medicine” in reality.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/depression

[2] Belmaker, R. H., & Agam, G. (2008). Major depressive disorder. New England Journal of Medicine, 358(1), 55-68.

[3] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Gocayne, J. D. (2001). The sequence of the human genome. science, 291(5507), 1304-1351.

[4] Boisvert, S., Laviolette, F., & Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Journal of computational biology, 17(11), 1519-1533.

[5] Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849.

[6] Cui, J., Chen, Y., Chou, W. C., Sun, L., Chen, L., Suo, J., ... & Kang, J. (2010). An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. Nucleic acids research, 39(4), 1197-1207.

[7] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3), 392-398.

[8] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

[9] Lee, I. H., Lushington, G. H., & Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. Journal of Clinical Bioinformatics, 1(1), 11.

[10] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3), 392-398.

[11] Boucheham, A., & Batouche, M. (2014, August). Robust biomarker discovery for cancer diagnosis based on meta-ensemble feature selection. In 2014 Science and Information Conference (pp. 452-560). IEEE.

[12] Hamet, P., & Tremblay, J. (2005). Genetics and genomics of depression. Metabolism, 54(5), 10-15.

[13] Wurtman, R. J. (2005). Genes, stress, and depression. Metabolism, 54(5), 16-19.

[14] https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE19738

[15] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

[16] Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators B: Chemical, 212, 353-363.

[17] Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson +Education Limited,.

[18] Karthik, S and Sudha, M. (2018), A Survey on Machine Learning Approaches in Gene Expression Classification in Modelling Computational Diagnostic System for Complex Diseases, International Journal of Engineering and Advanced Technology, 8:2, pp.182-199

[19] Sudha, M. (2017), Weather Modeling using Data-driven Adaptive Rough-Neuro-Fuzzy approach,

Current World Environment, 12: 02, pp. 429-435.
[20] Sudha, M. and Subbu, K. (2017).Statistical Feature Ranking and Fuzzy Supervised Learning Approach in Modeling Regional Rainfall Prediction Systems, AGRIS on-line Papers in Economics and Informatics, 09: 02, pp. 117-126
[21] Sudha, M. (2017). Intelligent decision support system based on rough set and fuzzy logic approach for efficacious precipitation forecast, Decision Science Letters, 06, pp. 96-105.
[22] Sudha, M. (2017), Instant Medical Care and Drug Suggestion Service using Data Mining and Machine Learning based Intelligent Self-Diagnosis Medical System, International Journal of Advanced Life Sciences, 10:3, p. 318
[23] Sudha, M. (2016). Disease diagnosis using association rule mining based knowledge inference system, International Journal on Pharmacy and Technology, 08: 03, pp. 16369-16379.
[24] Evolutionary and Neural Computing based Decision Support System for Disease Diagnosis from Clinical Data sets in Medical Practice, Springer Nature : Journal of Medical Systems, 41:178.
[25] Sudha, M. and B. Poorva (2019), B Predictive Tool for Dermatology Disease Diagnosis using Machine Learning Techniques, International Journal of Innovative Technology and Exploring Engineering.8:9.pp. 355- 451.
[26] Karthik, S., Perumal, R. S., & Mouli, P. C. (2018). Breast Cancer Classification Using Deep Neural Networks. In Knowledge Computing and Its Applications (pp. 227-241). Springer, Singapore.
[27] Sudha, M. (2017), Computational Intelligence based Sports Success Prediction System using Functional Pattern Growth Tree - A case study , International Journal of Computational Intelligence Research, 13:10, pp.2431-2438.