Diagnostic Gene Biomarker Selection for Alzheimer's Classification using Machine Learning

Karthik Sekaran, Sudha. M

Alzheimer's disease (AD), also referred to as Alzheimer's is a neurodegenerative disease and most common type of dementia. It starts at an older age and slowly progressive over time. It is a brain disease which causes loss of memory, reasoning and thinking capability of a person. Short-term memory loss is one of the main symptoms of the AD. Other common symptoms are said to be mood-swings, difficulty in understanding language and its interpretation etc. The major problem in the AD is, it can't be reverted, but controllable with proper treatment. Genetic factors have a high impact on developing an AD, which can be inherited through genes. According to recent studies, gene therapy shows better results for Alzheimer's patients than other common medications. It reduces the risk effect of the AD and has a gradual improvement on the patient's condition. So, identification of gene biomarkers, having high involvement in developing AD could improve positive response over the treatment. In this paper, gene expressions of AD patients and normal peoples are analyzed using statistical approaches and Machine Learning (ML) algorithms. Differential Gene Expression (DEG) identification has an important part in the selection of most informative genes. Potential gene biomarkers are selected using a meta-heuristic global optimization algorithm called Rhinoceros Search Algorithm (RSA). As an outcome from RSA, 24 novel gene biomarkers are identified. Four supervised ML algorithms such as Support Vector Machines (SVM), Random Forest (RF), Naïve Bayes (NB) and Multilayered Perceptron Neural Network (MLP-NN) are used for the classification of two different group of samples. Among them, RSA-MLP-NN model achieved 100% accuracy on identifying the distinction between AD and normal genes and proved its efficacy.

Index Terms*: *Alzheimer's disease, Biomarkers, Gene Expression, Healthcare, Machine Learning.

I. INTRODUCTION

The prevalence of Alzheimer's disease (AD) is increasing all over the world, with an increase in the population of aged peoples. Dementia is the main cause of developing an AD and it is said to be a neurodegenerative disease, with an approximate range of 50% - 70% cases [1, 2]. This disease has an irreversible effect on brain and can't be cured permanently but the progression can be restricted for some time with available treatment procedures [3]. A progressive decline in cognitive function in memory and behavior exposes the presence of AD. The estimated cost spent to treat AD is approximately $172 billion per year. Now currently there are 24 million people are diagnosed with dementia and it is expected to increase by four times by 2050 [4]. The increase in the risk factor of developing AD occurs at the age of 60 and older [5]. The causes of the AD are said to be aging, degeneration of anatomical pathways, environmental factors, malnutrition, head injuries, vascular factors, immune system dysfunction, infectious agents and genetic factors [6]. A recent research study reveals the strong association between AD and Cardio Vascular Disease (CVD) [7]. The symptoms of the AD are lack of understanding and interpreting language, short-term memory loss, cognitive impairment, problem-solving skills, confusion, difficulty in doing routine jobs and mood swings and learning ability [8]. Diagnosis of the AD is difficult and it is quiet dependent on functional and cognitive assessments [9]. Genetic factors are considered to be one of the reasons for the increase in risk factor of developing AD. Identifying potential gene biomarkers of the AD will pave a new way to find different therapies for effective treatment [10]. Around 60% to 80% chance is there that AD can be inherited [11]. This paper is focused on identifying novel gene biomarkers of the AD using computational techniques. The dataset used in this experiment is published by the National Center for Bioinformatics Information (NCBI) [12]. Gene expressions of a group of normal peoples and AD patients are analyzed with statistical measures. Top 100 DEG's are identified using t-statistics with the significance of p-value <0.05, which is assumed as informative genes. RSA is used for selecting the best feature subset. Different ML algorithms are deployed to classify the gene expression data of two different groups of samples. The following sections are discussed in detail in order. Section 2 discusses the background study of the related work. Section 3 covers the methods applied in these experiments and the materials used. Section 4 provides the experimental results achieved from this study. Section 5 discusses notable points about the experiment and section 6 concludes the work.

II. BACKGROUND

A genome-wide association study (GWAS) is performed on Alzheimer's gene expression dataset with brain-specific data. SVM algorithm is used to develop the predictive model. They achieved a ROC rate of 84.56% and 94% [20]. In another related work, MRI data is used to classify Alzheimer's patients with normal samples using machine learning methods.

Revised Manuscript Received on October 05, 2019.

Karthik Sekaran, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

Dr. M. Sudha, Associate Professor, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

Regularized logistic regression is used for feature selection and SVM is used for classification. ADNI database is used to perform this experimental study [21].

Gene expression signature identification from blood samples are made to predict Autism Spectrum Disorder (ASD). SVM, k-Nearest Neighbor and Linear Discriminant Analysis (LDA) is used for classification. Also, hierarchical clustering is performed to group the samples into different clusters [22]. A repetitive gene (CAPS2) is identified using ML which has a higher risk of developing ASD. Using Principal Component Analysis (PCA), the features are selected and SVM is used for classification [23].

A correlation-based feature selection algorithm is developed to identify the presence of schizophrenia. Locally Weighed Learning classifier delivered best results [24].

PRKCA is identified as a riskiest gene of developing Post-Traumatic Stress Disorder (PTSD), done by ML. Also, three other genes are identified which has a co-partial relation in developing PTSD. Cytoscape tool is used to analyze the pathway of gene expression and Partial Least Square is used for classification [25].

An analytical framework is developed to classify bipolar disorder with Significance Analysis of Microarrays (SAM). RFE-SVM is adopted to develop a predictive model. Two risk genes are identified namely NOG and CTBP1 [26]. A genetic risk score model is developed using the RF algorithm to classify between normal and abnormal genes [27].

A novel decision support system is developed using intelligent methods to identify best features for various ML algorithms to make better predictions from clinical information. A relative fitness function for genetic algorithm is calculated to find optimal features. For classification process, Back-Propagation NN is deployed [28].

Major Depressive Disorder gene expressions were analyzed using Weighed Multiple Logistic Regression techniques. Multiple Imputation techniques is used for noise removal process. In this work, no gene biomarkers are identified but classified the genes as normal and abnormal [29].

Alzheimer's gene expression analysis is carried out from 11 different data sources. The genes are initially weighed using various techniques such as SVM, information ratio and so on. Three different neural network algorithms are used to develop the predictive model [30].

III. MATERIALS AND METHODS

A.Microarray Data Acquisition

The dataset used in this experiment is taken from the publicly available data source namely Gene Expression Omnibus (GEO) maintained by NCBI. The accession number of the dataset is GSE1297 [13]. The data is extracted from the hippocampal region of the brain. Totally 31 samples are analyzed in this experiment. Among them, 9 samples are from control cases and 22 samples are from AD patients.

B.Data Preprocessing and DEG identification

The accessed dataset is in microarray data format as.CEL file. The probe values from the file are extracted using "Bioconductor" package from R Language. It supports many powerful libraries for microarray data processing such as limma, affy, GEOquery, and Biobase. Affy library is used to transform .CEL files to probe intensity matrix [14]. Three important steps such as background correction, normalization, and summarization of gene expression probe data are performed using Robust Multi-array Average (RMA) [15], median polish and quantile normalization techniques [16]. Significantly expressed genes are identified using the limma library using t-test [17]. Posterior data processing steps such as identification of regulated genes are calculated using log2 transformation on probe data. A significance of p-value<0.05 is framed to extract top 100 informative genes from a t-test. These genes are selected with respect to the smallest p-values on the top.

C.Rhinoceros Search Algorithm

Inspired by the natural behavior of rhinoceros, a meta-heuristic search algorithm called RSA is developed to solve optimization problems effectively [18]. In this work, to select the optimal feature subset from the top 100 informative genes, RSA is used. Some assumptions are presumed on developing this algorithm based on the rhinoceros behavior. Those are

- Male agents only have bouncing mechanism while females or not.
- Levy flight is done by all the agents in the population within the search range.
- The die-born mechanism is adopted at the rate of 5 percent of the population.

Algorithm : RS

Step 1: Input: $R_{ml:}$ Population of Male Rhinoceros; $R_{fl:}$ Population of Female Rhinoceros
Step 2: Define Size: $D_{ml:}$ Search range of Male rhinoceros (Radius); $D_{fl:}$ Search range of Female Rhinoceros (Radius)
Step 3: Output: global_best_value
Step 4: Parameter Initialization
Step 5: While (N < no_of_epoch & global_best_value not achieved) do:
For each male rhinoceros r_{ml} in R_{ml} do:
If Distance ($R_{ml, I}$, $R_{ml, j}$) < D_{ml}
Escape (R_{ml}.Weakest)
End-if
$R_{ml, N}$ = Levy_Flight($R_{ml, N-1}$)
Update_fitness_all_Male_rhinoceros(R_{ml})
End-for
For each female rhinoceros l_f in L_f do
$R_{fl,N}$ = Levy_Flight ($R_{fl,N-1}$)
Update_fitness_all_Female_rhinoceros(R_{fl})
End-for
For all rhinoceros (r_{ml} in R_{ml}) and (r_{fl} in R_{fl}) do
If life_test(r_k)=0, then
Remove_rhinoceros(r_k)
Born_Baby_rhinoceros(r_k)
End-for
If fitness (g_{best}) < fitness (global_best_value) then
global_best_value = g_{best} //Updating Global Best
End-if
N++
End-while

Table 1: Identification of Gene Biomarkers of Alzheimer's disease using RSA

ID	P.Value	Gene.symbol
215842_s_at	1.24E-05	ATP11A
222196_at	1.37E-05	LOC389906
217158_at	2.34E-05	PTGER4P2
212122_at	2.45E-05	RHOQ
211584_s_at	8.64E-05	NPAT
215410_at	0.000108	PMS2P9
201262_s_at	0.000147	BGN
215908_at	0.000208	
207034_s_at	0.000212	GLI2
213205_s_at	0.000279	RAD54L2
205296_at	0.000316	RBL1
221729_at	0.000413	COL5A2
211496_s_at	0.000528	PDC
214177_s_at	0.000707	PBXIP1
203195_s_at	0.000738	NUP98
213981_at	0.000776	COMT
218551_at	0.000874	MIIP
207440_at	0.001152	SLC35A2
207961_x_at	0.001187	MYH11
220394_at	0.001215	FGF20
200928_s_at	0.001262	RAB14
200793_s_at	0.001301	ACO2
209643_s_at	0.001344	PLD2
203894_at	0.001487	TUBG2

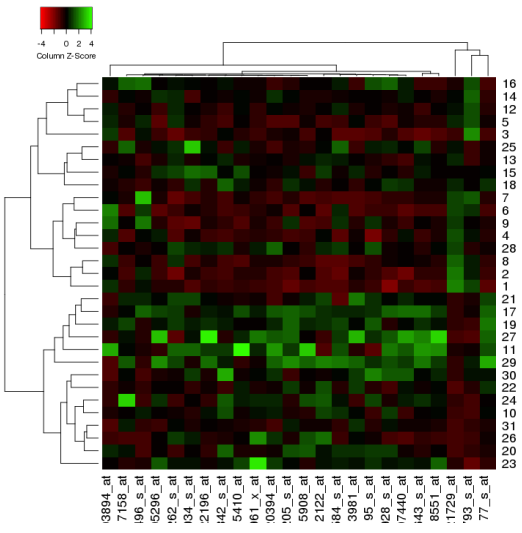

Fig 2: Heat map Expression of the data after RSA

D.Supervised Model for Alzheimer's Classification

After the identification of Alzheimer's risk genes from RSA, four ML algorithms are deployed to develop predictive models. SVM, RF, NB, and MLP-NN are the algorithms classify the data. Each model is trained and tested with the 31 samples contains 9 control and 22 AD patients. The results of each model are defined in Table 2 and Table 3. From Table 3, it clearly states that RSA-MLP-NN achieved the best result with 100% classification accuracy than other models. K-Fold cross-validation is applied on the dataset to estimate the performance of the predictive model with 10-folds each.

Table 2: Performance of ML Algorithms with top-100 genes (before RSA)

	Acc (%)	Sen (%)	Spec(%)
SVM	93.55	88.89	95.45
NB	90.32	87.50	91.30
RF	90.32	100	88
MLP-NN	96.77	90	100

Table 3: Performance of ML Algorithms (after RSA)

	Acc (%)	Sen (%)	Spe(%)
SVM	87.10	85.71	87.50
NB	90.32	87.50	91.30
RF	96.77	100	95.65
MLP-NN	100	100	100

IV. RESULTS AND DISCUSSION

In this experiment, initially top 100 informative genes are selected from t-test with the significance of p-value <0.05. Among the 100 genes, to select optimal gene biomarkers, RSA is applied. Twenty four novel gene biomarkers are selected as an outcome of RSA. A clear description of those genes is clearly stated in Table 1. Four of them are down regulated and remaining 20 genes are up-regulated. Then, to classify the data of two different groups, SVM, NB, RF, and MLP-NN are applied to the reduced dataset. These classifiers give 87.10%, 90.32%, 97.66% and 100% accuracy respectively. By analyzing the results obtained, it is clear that RSA-MLP-NN model outperformed other benchmarking classifiers. Moreover, the results obtained from this study is compared with other related works are depicted in Table 4.

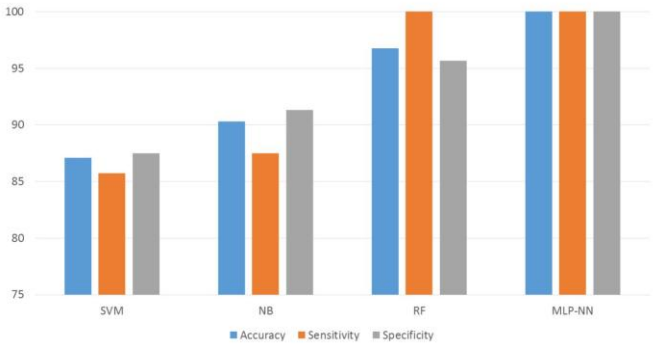

Fig 3: Comparison of performance of four ML algorithms (after RSA Feature Selection)

The objective of this study is to identify potential gene biomarkers of Alzheimer's disease by analyzing the microarray gene expression data of both normal and AD patients.

To achieve that, statistical and machine learning techniques are adopted and tested in the experimental dataset. Top DEG's are selected from the entire gene probes using t-test with a statistical significance of p-value <0.05. RSA is used to extract potential biomarkers from top 100 genes. Our findings highlight the presence of 24 important gene biomarkers on AD patients. It gives a clear distinction between healthy peoples and AD patients. To classify between two different groups, SVM, RF, NB, and MLP-NN are used. Among them, RSA-MLP-NN model achieved the best result and outperformed the other benchmarking models with RSA. The results are plotted as a bar graph in Figure 2 to visually compare and analyze. These results are also compared with the existing related works in order to evaluate the performance of this work, given briefly in Table 4.

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.

V. CONCLUSION

In this experimental work, 31 gene expression samples of normal and AD patients are analyzed to identify the risk gene biomarkers of Alzheimer's disease. This study finds 24 novel risk genes, provides the distinction between normal and disease states. Machine learning algorithms are used to classify the groups. This study revealed that RSA-MLP-NN model performs well than other benchmarked models compared in this study. The main limitation of this experiment is the sample size. The frequency of generating gene expression data is often very less when compared with medical records. So, in this study, with the available data, a better model is developed to identify AD risk genes. With the advent of advanced techniques in bioinformatics, the generation of the genetic data will be more. In future, complex genetic structures will be analyzed and decoded using computational techniques and intelligent algorithms with more to provide personalized treatment like gene therapy for every patient based on the uniqueness in their genes for responsive treatment with effective medication strategies.

REFERENCES

1. Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., Cedazo-Minguez, A., Dubois, B., Edvardsson, D., Feldman, H. and Fratiglioni, L., 2016. Defeating Alzheimer's disease and other dementias: a priority for European science and society. The Lancet Neurology, 15(5), pp.455-532.
2. Kumar, A. and Singh, A., 2015. A review of Alzheimer's disease pathophysiology and its management: an update. Pharmacological Reports, 67(2), pp.195-203.
3. Liu, M., Cheng, D. and Yan, W., 2018. Classification of Alzheimer's Disease by Combination of Convolutional and Recurrent Neural Networks Using FDG-PET images. Frontiers in Neuroinformatics, 12, p.35.
4. Reitz, C. and Mayeux, R., 2014. Alzheimer disease: epidemiology, diagnostic criteria, risk factors, and biomarkers. Biochemical Pharmacology, 88(4), pp.640-651.
5. Fjell, A.M., McEvoy, L., Holland, D., Dale, A.M., Walhovd, K.B. and Alzheimer's Disease Neuroimaging Initiative, 2014. What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus. Progress in neurobiology, 117, pp.20-40.
6. Armstrong, R.A., 2013. What causes Alzheimer's disease?. Folia Neuropathologica, 51(3), pp.169-188.
7. de Bruijn, R.F. and Ikram, M.A., 2014. Cardiovascular risk factors and future risk of Alzheimer's disease. BMC medicine, 12(1), p.130.
8. Alzheimer's Association, 2018. 2018 Alzheimer's disease facts and figures. Alzheimer's & Dementia, 14(3), pp.367-429.
9. Marcello, E., Gardoni, F. and Di Luca, M., 2015. Alzheimer's disease and modern lifestyle: what is the role of stress?. Journal of neurochemistry, 134(5), pp.795-798.
10. Bekris, L.M., Yu, C.E., Bird, T.D. and Tsuang, D.W., 2010. Genetics of Alzheimer disease. Journal of geriatric psychiatry and neurology, 23(4), pp.213-227.
11. Rongve, A., Årsland, D. and Graff, C., 2013. Alzheimer's disease and genetics. Tidsskrift for den Norske laegeforening: tidsskrift for praktisk medicin, ny raekke, 133(14), pp.1449-1452.
12. Edgar, R., Domrachev, M. and Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), pp.207-210. (GEO)
13. Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R. and Landfield, P.W., 2004. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proceedings of the National Academy of Sciences, 101(7), pp.2173-2178.
14. Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A., 2004. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 20(3), pp.307-315.
15. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2), pp.249-264.
16. Bolstad, B.M., Irizarry, R.A., Åstrand, M. and Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2), pp.185-193.
17. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), pp.e47-e47. (limma)
18. Tian, Z., Fong, S., Tang, R., Deb, S. and Wong, R., 2016, November. Rhinoceros Search Algorithm. In Soft Computing & Machine Intelligence (ISCMI), 2016 3rd International Conference on (pp. 18-22). IEEE.
19. Paylakhi, S., Paylakhi, S.Z. and Ozgoli, S., 2016. Identification of Alzheimer disease-relevant genes using a novel hybrid method. Progress in Biological Sciences, 6(1), pp.37-46.
20. Huang, X., Liu, H., Li, X., Guan, L., Li, J., Tellier, L.C.A.M., Yang, H., Wang, J. and Zhang, J., 2018. Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning. BMC neurology, 18(1), p.5.
21. Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J. and Alzheimer's Disease Neuroimaging Initiative, 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage, 104, pp.398-412.
22. 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.
23. Hameed, S.S., Hassan, R. and Muhammad, F.F., 2017. Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm. PloS one, 12(11), p.e0187371.
24. Zhang, H., Xie, Z., Yang, Y., Zhao, Y., Zhang, B. and Fang, J., 2017. The correlation-base-selection algorithm for diagnostic schizophrenia based on blood-based gene expression signatures. BioMed Research International, 2017.
25. Dong, K., Zhang, F., Zhu, W., Wang, Z. and Wang, G., 2014. Partial least squares based gene expression analysis in posttraumatic stress disorder. Eur Rev Med Pharmacol Sci, 18(16), pp.2306-2310.

26. Leska, V., Bei, E.S., Petrakis, E. and Zervakis, M., 2016. Gene Expression Data Analysis for Classification of Bipolar Disorders. In XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016 (pp. 500-506). Springer, Cham.
27. Chuang, L.C. and Kuo, P.H., 2017. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. Scientific Reports, 7, p.39943.
28. Sudha, M., 2017. Evolutionary and Neural Computing Based Decision Support System for Disease Diagnosis from Clinical Data Sets in Medical Practice. Journal of medical systems, 41(11), p.178.
29. Dipnall, J.F., Pasco, J.A., Berk, M., Williams, L.J., Dodd, S., Jacka, F.N. and Meyer, D., 2016. Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. PloS one, 11(2), p.e0148195.
30. Barati, M. and Ebrahimi, M., 2016. Identification of Genes Involved in the Early Stages of Alzheimer Disease Using a Neural Network Algorithm. Gene, Cell and Tissue, 3(3).
31. 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
32. Sudha, M. (2017), Weather Modeling using Data-driven Adaptive Rough-Neuro-Fuzzy approach, Current World Environment, 12: 02, pp. 429-435.
33. 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
34. 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.
35. 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
36. Sudha, M. (2016). Disease diagnosis using association rule mining based knowledge inference system, International Journal on Pharmacy and Technology, 08: 03, pp. 16369-16379.
37. 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.
38. 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.
39. 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.
40. 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.

AUTHORS PROFILE

Mr. Karthik Sekaran received M. Tech in Software Engineering from Vellore Institute of Technology, Vellore, India in 2017. He is currently full time Research Scholar pursuing PhD in Information Technology and Engineering from VIT. His current research interests are in Bioinformatics, Machine Learning and Computational Intelligence.

Dr. M. Sudha, is currently working as Associate Professor in the department of Information Technology of School of Information Technology and Engineering at VIT. She has 17 years of teaching cum research experience. Her research expertize are in the field of Medical Informatics, Machine Learning and Ambient Intelligent Computing. She is a member of Association of Computing Machinery (MACM).