Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15327 - 15330
Received 05 March 2021; Accepted 01 April 2021.

A Computational approach for the Identification of Parkinson's Disease Gene

Rhea Mary Josi [1], Dr R.I. Minu [2]
Pg Student [1*], Asso.Prof [2], [1,2] Department Of Computer Science,
[1,2] Srm Institute Of Science And Technology, Chennai, India
[*] corresponding author: rj2939@srmist.edu.in

ABSTRACT:
Parkinson's disease is a serious brain disorder which affects the nervous system and coordination. Genetic and environmental factors might have an involvement in the causing of this syndrome but the true causes and the symptoms are still not predictable. Therefore, early detection and diagnosis can be done only using the identification of disease genes. These disease gene identifications can be done using different computational approaches. The proposed system is a prediction model to detect Parkinson's disease gene from the normal genes using isolation forest algorithm.
Keywords: Parkinson's disease, dopamine, Alzheimer's disease, Isolation forest algorithm, anomaly detection

INTRODUCTION

Parkinson' s disease is one of the worst serious illness or neurological condition [9]. Alzheimer's and Parkinson's disease are the 1st and 2nd serious and worst neurological condition. It mainly affects the entire nervous system of humans and adversely affects the movement and coordination of an individual. The initial symptoms of this syndrome can be shakes or tremors since substantia nigra is the region of the brain that gets affected. Substantia nigra is the area which controls the balance and movement [27]. It is a slow progressing disorder where it starts affecting a part of the body especially on a limb. Tremors slowly affects the arm, legs and feet. The tissues in the substantia nigra stops producing a chemical messenger called dopamine for the people affected with this disease. As dopamine reduces, an individual loses the ability to talk, coordinate and move. This brain disorder can worsen up over the time.

This disease does not have a permanent cure but medicines, surgical treatments and other therapies can help in reducing the effect of the symptoms. There are no perfect diagnostic mechanisms like blood tests or laboratory tests which diagnose the Parkinson's disease. It can only be understood with the patient's medical history and records. The similar symptoms can occur for other diseases also, so why and how it happens and the causes of the disorder is still unknown. Therefore, some patients might misdiagnose with this disorder.

Parkinson's disease can occur to both males and females. But researches prove that men are likely to get affected with the disease when compared to women [24]. This is mostly found in individuals after 50 years of age and is called a late onset disease. When the symptoms and signs occur for persons below 50 years of age is considered to be very rare and is called the early onset of disease. How mutated genes cause Parkinson's disease or develop the symptoms is not completely known [27]. Researches reveals that cases of Parkinson's disease caused by genetic mutations are rare. 15% of the persons suffering from this disorder only show the history of this disorder. Genes which do not possess this disorder even gets mutated, modified and can thus be dangerous.

The disorder like Parkinson's disease, Alzheimer's disease is not easily treatable and will result in the permanent illness or demise of the affected people. Thus, identification and detection of disease-causing genes are very essential and necessary which can help in the early diagnosis of these diseases. Computational approaches can be used for the identification of detection of genes. Computational approach can be better when compared to the biological methods as it is time consuming and computational costs are high. Many machine learning approaches are used in the gene identification and detection of these types of spasmodic disorders. R, python, Perl are the programming languages which can be used for the pre-processing and analysis of all kinds of genetic datasets.

BACKGROUND THEORY

Several computational approaches are developed to predict the disease genes using various ML algorithms. Isolation forest algorithm is an anomaly detection technique used in the proposed system for the prediction model.

Isolation forest is an unsupervised learning mechanism which works by the principle of anomaly isolation. The word 'isolation' is defined as 'distinguishing an instance from the rest of the instances [25]. Anomalies are an observation or an instance that deviates it's characteristics from other instances. Isolation forest is a varied approach when compared to the other algorithms as it isolates the anomalous ones from the datasets. A dataset containing PD genes can be trained using this algorithm and then the normal genes whenever encountered can be detected as an anomaly. The principle of this algorithm is to build a forest (iForest) containing iTrees [26]. A training data with numerical data types, the data is recursively divided until iTree differentiates each data from other data [26].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15327 - 15330
Received 05 March 2021; Accepted 01 April 2021.

IMPLEMENTATION

The proposed system consists of 3 modules:
1. Data pre-processing
2. Dimensionality Reduction
3. Isolation forest algorithm

Data pre-processing and dimensionality reduction are done to a dataset containing both normal and disease genes. Data pre-processing is done to clean the acquired dataset of genes. It involves handling missing values, encoding categorical data etc, Dimensionality reduction is the process to reduce the higher dimensions into lower dimensions. Since gene datasets are distorted and large dimensionality reduction is very important. Then for the third module isolation forest algorithm is used for the anomaly detection.

Dataset Availability:

The dataset has been collected from a website named 'Gemma' which is used for the analysis, sharing and using of genomic data.

https://gemma.msl.ubc.ca/expressionExperiment/showExpressionExperiment.html?id=529

The top 250 PD genes used is a dataset which is derived from the above given dataset. It is obtained from 'Kaggle'.

https://www.kaggle.com/andrewgao/top-250-parkinsons-genes

STEPS

Data Pre-processing:

1. Getting the dataset
2. Importing libraries
3. Finding and handling missing data
4. Encoding categorical data
5. Splitting of training and testing of data.

Dimensionality Reduction:

1. Choose the size for reducing the original features
2. Define encoding layer
3. Define decoding layer
4. Build the autoencoder model

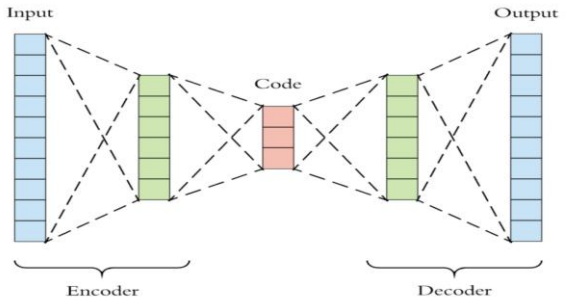

Fig 1: Autoencoder model

Isolation Forest Algorithm:

This is done to a dataset containing PD genes, and these are trained in such a way that when a normal gene is encountered it is treated as an anomaly.

1. Importing the libraries
2. Contamination is defined.
3. Training the dataset
4. Anomaly detection

RESULTS AND DISCUSSION

The Parkinson's disease genes are taken from the dataset and the 250 PD genes (approx..) are fed to the anomaly detection algorithm and are trained. Once a gene other than the PD genes are fed to the algorithm then an anomaly is detected. When the PD genes are encountered then it shows a 1 as output and if a normal gene or a gene other than the PD genes are encountered then the output is -1.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15327 - 15330
Received 05 March 2021; Accepted 01 April 2021.

```
In [4]:  from sklearn.preprocessing import LabelEncoder
         import pickle

         le = LabelEncoder()
         for i in df:
             if df[i].dtypes == 'object':
                 df[i] = le.fit_transform(df[i])
                 filename = 'LE/{}.sav'.format(i)
                 pickle.dump(le, open(filename, 'wb'))
         df.head()

Out[4]:
```

	A	B	C	D	E	F	G	H
0	21	0	7	93	6	97	60	107
1	224	3	95	60	89	96	164	179
2	135	0	6	94	5	95	138	156
3	113	0	40	81	39	94	91	146
4	188	0	10	91	9	93	149	160

Fig 2: Label encoding of the dataset

```
In [5]:  from sklearn.preprocessing import MinMaxScaler

         min_max_scaler = MinMaxScaler()
         x = min_max_scaler.fit_transform(df)
         x

Out[5]:  array([[0.08974359, 0.        , 0.02991453, ..., 0.41452991, 0.2690583 ,
                 0.47982063],
                [0.95726496, 0.33333333, 0.40598291, ..., 0.41025641, 0.73542601,
                 0.80269058],
                [0.57692308, 0.        , 0.02564103, ..., 0.40598291, 0.61883408,
                 0.69955157],
                ...,
                [0.47435897, 0.        , 0.16239316, ..., 0.99145299, 0.3632287 ,
                 0.5470852 ],
                [0.74786325, 0.        , 0.05128205, ..., 0.9957265 , 0.70403587,
                 0.77578475],
                [1.        , 1.        , 1.        , ..., 1.        , 0.38116592,
                 0.12107623]])

In [6]:  from sklearn.ensemble import IsolationForest

         clf = IsolationForest(n_estimators=100, contamination=0.0,random_state=42)
         clf.fit(x)

Out[6]:  IsolationForest(contamination=0.0, random_state=42)
```

Fig 3: Normalization of the dataset.

```
In [6]:  from sklearn.ensemble import IsolationForest

         clf = IsolationForest(n_estimators=100, contamination=0.0,random_state=42)
         clf.fit(x)

Out[6]:  IsolationForest(contamination=0.0, random_state=42)

In [9]:  filename = 'model.sav'
         pickle.dump(clf, open(filename, 'wb'))

In [10]: y = clf.predict(x)
         y

Out[10]: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
```

Fig 4: Output of the model encountering PD gene (output = 1).

CONCLUSION

This model is used for disease gene prediction of Parkinson's disease. The anomaly detection algorithm isolation forest is used in the model which can be efficiently used in datasets with low time complexity.

FUTURE WORK

It can be done to very large datasets. Future research can be done in other disease genes also.

REFERENCES

[1] Peng, J., Guan, J., & Shang, X. (2019). Predicting Parkinson's disease genes based on node2vec and autoencoder. *Frontiers in genetics*, *10*, 226.

[2] Li, Y. I., Wong, G., Humphrey, J., & Raj, T. (2019). Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nature communications*, *10*(1), 1-10.

[3] El Haddad, S., Serrano, A., Moal, F., Normand, T., Robin, C., Charpentier, S., & Ozsancak, C. (2020). Disturbed expression of autophagy genes in blood of Parkinson's disease patients. *Gene*, *738*, 144454.
[4]. George, G., Singh, S., Lokappa, S. B., & Varkey, J. (2019). Gene co-expression network analysis for identifying genetic markers in Parkinson's disease-a three-way comparative approach. *Genomics*, *111*(4), 819-830.
[5]. Senturk, Z. K. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. *Medical Hypotheses*, *138*, 109603.
[6]. Krishnagopal, S., Coelln, R. V., Shulman, L. M., & Girvan, M. (2020). Identifying and predicting Parkinson's disease subtypes through trajectory clustering via bipartite networks. *PloS one*, *15*(6), e0233296.
[7]. Rastegar, D. A., Ho, N., Halliday, G. M., & Dzamko, N. (2019). Parkinson's progression prediction using machine learning and serum cytokines. *NPJ Parkinson's disease*, *5*(1), 1-8.
[8]. Sadek, R. M., Mohammed, S. A., Abunbehan, A. R. K., Ghattas, A. K. H. A., Badawi, M. R., Mortaja, M. N., ... & Abu-Naser, S. S. (2019). Parkinson's Disease Prediction Using Artificial Neural Network.
[9]. S. Bhat, U.R. Acharya, Y. Hagiwara, N. Dadmehr, H. Adeli, Parkinson's disease: Cause factors, measurable indicators, and early diagnosis, Computers in Biology and Medicine (2018)
[10]. Le, D. H. (2020). Machine learning-based approaches for disease gene prediction. *Briefings in functional genomics*.
[11]. Xu, L., Liang, G., Liao, C., Chen, G. D., & Chang, C. C. (2018). An efficient classifier for Alzheimer's disease genes identification. *Molecules*, *23*(12), 3140.
[12]. Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics*, *15*(1), 41-51.
[13]. Almugren, N., & Alshamlan, H. M. (2019). New bio-marker gene discovery algorithms for cancer gene expression profile. *IEEE Access*, *7*, 136907-136913.
[14]. Vinitha, S., Sweetlin, S., Vinusha, H., & Sajini, S. (2018). Disease prediction using machine learning over big data. *Computer Science & Engineering: An International Journal (CSEIJ)*, *8*(1).
[15]. Asif, M., Martiniano, H. F., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. *PloS one*, *13*(12), e0208626.
[16]. Zhou, H., & Skolnick, J. (2016). A knowledge-based approach for predicting gene–disease associations. *Bioinformatics*, *32*(18), 2831-2838.
[17]. Yang, K., Wang, R., Liu, G., Shu, Z., Wang, N., Zhang, R., ... & Zhou, X. (2018). HerGePred: heterogeneous network embedding representation for disease gene prediction. *IEEE journal of biomedical and health informatics*, *23*(4), 1805-1815.
[18]. Ata, S. K., Ou-Yang, L., Fang, Y., Kwoh, C. K., Wu, M., & Li, X. L. (2018). Integrating node embeddings and biological annotations for genes to predict disease-gene associations. *BMC systems biology*, *12*(9), 138.
[19]. Picart-Armada, S., Barrett, S. J., Willé, D. R., Perera-Lluna, A., Gutteridge, A., & Dessailly, B. H. (2019). Benchmarking network propagation methods for disease gene identification. *PLoS computational biology*, *15*(9), e1007276.
[20] Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864).
[21] Won, J. H., Kim, M., Youn, J., & Park, H. (2020). prediction of age at onset in parkinson's disease using objective specific neuroimaging genetics based on a sparse canonical correlation analysis. *Scientific Reports*, *10*(1), 1-12.
[22] Nikdelfaz, O., & Jalili, S. (2018). Disease genes prediction by HMM based PU-learning using gene expression profiles. *Journal of biomedical informatics*, *81*, 102-111.
[23] Davie, C. A. (2008). A review of Parkinson's disease. *British medical bulletin*, *86*(1), 109-127.
[24] Radhakrishnan, D. M., & Goyal, V. (2018). Parkinson's disease: A review. *Neurology India*, *66*(7), 26.
[25] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.
[26] Xu, D., Wang, Y., Meng, Y., & Zhang, Z. (2017, December). An improved data anomaly detection method based on isolation forest. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 2, pp. 287-291). IEEE.
[27] JOSI, R. M., & MINU, R. (2020, December). Review of computational approaches to Parkinson's disease gene prediction. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 780-788). IEEE.