SN Computer Science (2020) 1:118
https://doi.org/10.1007/s42979-020-0099-4

ORIGINAL RESEARCH

Machine Learning Techniques to Identify Dementia

Nivedita Manohar Mathkunti[1] · **Shanta Rangaswamy**[1]

Abstract
The challenge to clinician is to diagnosis the complex disease such as Alzheimer, disease, frontotemporal dementia and Parkinson. The diseases are bit complex to diagnose in terms of symptoms as they overlap in many aspects. So, it is necessary to investigate the process of diagnostic with more accuracy with different parameters of the disease. The paper presents the implementation of machine learning algorithms to get more accuracy to identify the disease Parkinson. The data set referred is from Online-based machine learning repository also recognized as UCI. Support vector machine, *K*-nearest neighbor and linear discriminant analysis algorithms are used to calculate accuracy, recall and confusion matrix. The outcome of this implementation has given the accuracy of 100% for SVM and KNN algorithms and 80% for LDA. The size of the data set is 196 entries, and a number of parameters for each row are 22. The main parameters which are significant for the disease are 5, and implementation has done with 2 different versions. In first version, all the parameters are included, and in the second version, only 5 main parameters included. In both cases, training set is 70% and test set is 30%. SVM shows more accuracy in both versions of implementation. This helps to find the diagnosis and also to judge the significance of parameters with real data set.

Keywords Dementia · *K*-Nearest neighbor algorithms · Linear discriminant analysis · Machine learning algorithms · Parkinson · Support vector machine

Introduction

All over the world challenges of diagnosis of complex diseases marked as a one of the important investigations toward the building of healthy society. In health care, global burden is to investigate and to get accurate diagnosis process for neurological disorders. According to World Health Organization (WHO) report, it is predicted that neurological disorder may reach 6.77% by 2030. The high mortality because of neurological disorder [1] is also reported. The other brain disease, frontotemporal dementia (FTD), is clinically as well as a genetically syndrome of heterogeneous, which is featured by overlying clinical symptoms. These symptoms involve changes in behavior, language motor function, personality and degeneration of the brain region, frontal and temporal lobes [2]. FTD causes the cognitive and social decline of relative to the adult level of capability. The range of age of FTD patients is 40–82 years. The major phenotypes of this disease are a disorder of social activity and execution of any functioning. This rate is around 55% and similarly for primary progressive aphasia (PPA) which rate around 45% [3]. PPA is main type of neurological syndrome leading to language impaired gradually. Based on the parts of the hemisphere of left side and with other types of aphasia, PPA symptoms are accompanied with spoiling sensitivity.

However, unlike, disparate most other aphasias, PPA results from a continuous deterioration in brain tissue, which leads to early symptoms being far less damaging than later symptoms. Motor disorder of the brain is mainly due to progressive supranuclear palsy (PSP). It is known as the Steele–Richardson–Olszewski syndrome. It is also a degenerative disease with involvement of gradual deterioration. In some cases, the death/loss of some specific volumes of the brain causes this disease. At present for every 100,000 population, there will be at least 6 patients of PSP and is described as a tauopathy and

This article is part of the topical collection "Advances in Computational Intelligence, Paradigms and Applications" guest edited by Young Lee and S. Meenakshi Sundaram.

✉ Nivedita Manohar Mathkunti
niveditagg@gmail.com

[1] Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bengaluru, India

Published online: 10 April 2020

corticobasal syndrome (CBS) [4]. FTD also overlaps with a disease such as AD which may at the age range of 65 years above. The diagnosis of such death sentence disease is very much required at least to stretch the survival of the patient. The umbrella of the artificial intelligence (AI) has a subgroup called ML, and for the healthcare, ML is the future.

The work is done diagnosis of other disease Parkinson. Parkinson's disease (PD) is a nervous system's degenerateness. Because of this, motor system of the body affects. The weakness of voice continues in 90% of the patients of the age over 50 years [5–7]. It is observed that PD patients will augment with the aging of the worldwide population and deteriorate with time. This disease is most expensive disease to diagnose as well it leads to errors. Such diseases are to be diagnosed in early stage to avoid the end life.

In the current paper, the work is done on UCI data of PD. The implementation gave information of accuracy for the different algorithms in the way of support vector machine (SVM), *K*-nearest neighbor (KNN) and linear discriminant analysis (LDA). The paper proposes these algorithms to diagnose other disease FTD, Alzheimer with the different ML algorithms. "Literature Survey" section is about literature survey of both diseases and ML to diagnose. "Methodology", "Results and Analysis" and "Conclusion" sections reveal the methodology and results analysis and conclusion and future work, respectively. Most of the references for this paper are from PubMed, Springer, Nature, Elsevier, Brain and Translational neurodegeneration with the keyword as frontotemporal dementia, progressive non-fluent aphasias, progressive supranuclear palsy (PSP), FTD magnetic resonance image (MRI), ML in diagnosis, ML in FTD, etc. Some of the conference video lecturers of dominant doctors are referred.

The paper sections are as follows: "Literature Survey" section is about literature survey, in which references are from reputed journals such as PubMed, Springer, Elsevier, Wiley, Journal papers, Brain and also IEEE. "Methodology" section is about methodology which specifies about the algorithms used, "Results and Analysis" section is about results and analysis, and it shows accuracy, precision and recall for SVM, LDA and KNN algorithms in both training and test data set implementation separately of the referred data. The ratio of train and test data set is 70% and 30%. The data set includes 22 parameter and total 194 data. "Conclusion" section elucidates about the inference and also about future work.

Literature Survey

The paper by team of Tan et al. was intended to inspect about the connotation of decline in the hemoglobin levels with indications of cerebral small vessel disease (CSVD), and decline in the cognitiveness in an aged populace. The entire set of 796 data of different countries of age group 60 years was considered by them. They concluded that variation in levels of hemoglobin was accompanied with lobar microbleeds, cognitive, cognitive dysfunction and neurodegenerative signs [8]. The team of Ismai et al. worked with an objective to find the volume of hippocampal of adult's both male and female. The data set of 60 patients was used in which 28 were males and 32 were females. With the help of MRI-based volumetric measurement of the hippocampus region of the brain, they found a significant negative correlation between hippocampus volume and age with value *pf* as 0.017 and $r = 0.449$ which showed the significant correlation between the age and the modified hippocampus volume with $p = 0.386$ and $r = 0.170$ [9]. Team of Lee et al. worked with an objective to determine frontal lobe volume and temporal cortex thickness. They also have an objective to get the association between onset of AD with psychosis. To achieve this objective, they had taken 3-T magnetic resonance imaging and 3D magnetization of all 26 AD patients with psychosis and 48 AD patients without psychosis. Thickness of cortical and volume was measured, with the help of ANOVA measured differences. They concluded that volume of hippocampal was smaller in AD+P compared to AD−P and also found that right hippocampal atrophy is independently associated with AD+P [10, 11]. Lau and his team worked to investigate the most critical problem of white matter hyperintensities (WMH) using magnetic resonance images (MRI) with machine learning approach. The data size of 180 communal dwelling, dementia free, stroke-free healthy subjects was considered and performed automatic retinal image analysis. This was done by acquiring non-mydriatic retinal fundus images. The upshot of the investigation was the significant amount of WMH on brain with grade ≥ 2. Finally, they concluded 59% hypertension, 26% hyperlipidemia and severe WMH of 36%. The correlation coefficient was of 0.897 [12].

Team of McCarthy et al. intended to assess the state of literature with use of morphometric MRI to distinguish frontotemporal dementia (FTD) and judge the pertinence for clinical practice. They studied the usage of machine learning algorithms such as random forest to measure cortical thickness, logistic regression to measure volumes, SVM, naïve Bayes and leave-one-out cross-validation (LOOCV) to investigate VBM-GM density. They also concluded potentiality of morphometric MRI and its capability to improve the diagnostic methods of complex brain diseases [13]. Team of Garrard et al. worked to diagnose and laterality consequence in semantic dementia. The objective of the research was to diagnose semantic dementia and to find its effects. The methodology applied is naïve Bayes multinomial algorithms. Information gain also computed to identify the vocabulary

features to distinguish both healthy and semantic dementia patients. The research resulted with high accuracy for the applied algorithm [14].

Methodology

The important objective of machine learning algorithms is explanation and prediction. The thinks can be predicted by use of the current variable in the available database; prediction is possible by use of existing variables itself in pursuance to predict strange or future values of unknown or future values of interest [14]. The algorithms such as SVM, LDA and KNN are implemented to study the accuracy level of the disease.

Support Vector Machine

The SVM is one which is the most widespread algorithms in machine learning, and it is an outstanding procedure to make trial. The main three characteristics of SVM are—constructing a margin separator to maximize the decision boundary with more possible distance. This feature of SVM helps to generalize. SVM also generates a linear extrication hyperplane, but they have the caliber to implant the data addicted to a high-dimensional space, using kernel trick. SVMs are a nonparametric procedure that they recall training examples and potentially need to store them all. In a supervised learning problem, a data set will be considered which will be of finite set of real vectors with n features each which can be expressed as in Eq. 1. The algorithms used are SVM, LDA and KNN. The data set of Parkinson is taken from UCI. The data include a total of 22 subjects, and among them, 5 are main parameters which are significant for Parkinson. To diagnose the disease with machine learning algorithm with high accuracy, in the current research above-mentioned algorithms are used. Each algorithms result is not same in terms of recall, precision and accuracy.

$$X = \left\{\overline{X_1}, \overline{X_2}, \overline{X_3}, , , , , \overline{X_n}\right\} \quad \text{Where} \quad \overline{X_i} \in R_m \tag{1}$$

Practically, SVMs frequently end up with retentive mode with tiny fractions of the number of examples. In this way, SVM mix the cons of all the nonparametric and parametric models. This algorithm has the elasticity to represent multifaceted functions to denote complex functions, but they are resilient to overfitting [15, 16].

The SVM will be implemented in three sections as initialization, training and classification. In initialization, compute the distances between the datapoints and parameters of Kernel. In training, recognize the vector which are support vectors within some specified length/distance to the nearby points and position of the training data. The classification phase of an algorithm calculates the inner product of the test data and support vectors and accomplishes the classification [17].

Linear Discriminant Analysis

LDA is one of the common techniques which can be applied for a dimensionality reduction difficulty as a preprocessing phase for the machine learning and pattern classification applications data mining, biometric, bioinformatics and information retrieval.

In all these, application goals are to diminish the dimensions by eradicating the redundant and reliant on features by transforming the features from a higher-dimensional space that may lead to a cure of dimensionality problem to a space with lower dimensions. Table 2 shows the LDA algorithm in class-independent manner. A lower space of dimension of an independent class is calculated for all classes. A transformation of a matrix is computed for all classes are anticipated on selected eigenvectors.

The concept of LDA is similar to another algorithm PCA, but LDA is a superior feature extraction technique for classification tasks. In LDA, data are normally distributed and classes will have similar covariance matrices and that the features are statistically independent of each other, get the input as a N sample representing as a matrix. Calculate the means of all classes and also calculate the total mean of all data calculate matrix of class SB ($M \times M$) calculate matrix of class (within) of each class SW ($M \times M$). From the computed SB and SW matrix, W that maximize Fisher's formula and eigenvalues (λ) and eigenvectors (V) of W are computed. According to the equivalent eigenvalues, sort eigenvectors in descending order. Lower-dimensional space (Vk) of eigenvectors is taken from first K of eigenvectors. Original samples of (X) are to be projected onto the lower-dimensional space of LDA [17].

K-Nearest Neighbor Algorithms

The KNN algorithm is one of simple learning algorithm, and it relies on the hypothesis that "things that look alike." In present research, Minkowski distance formula is used. In universal, to predict qualitative responses, Bayes classifier will be used. With the real-time data, Bayer classifier is impossible because it cannot be known the conditional distribution of Y given X and also to classify a given comment to the class with advanced assessed probability. For such scenario, KNN is used which is acceptable for the given integer K and observation of test X_0. At the beginning, KNN identifies the K point in the train set of closet data set to X_0. This represented N_0 the probability for class j as part of points in N_0 and its response values equivalent to j as given in Eq. 2. At the end of KNN, it also applies Bayes rule and

Table 1 Confusion matrix of test data using KNN algorithm with all features

True label	Predicted label	
	Negative	Positive
Negative	TN = 9	FP = 6
Positive	FN = 0	TP = 44

classifies the observation X_0 to the class mentioned with probability [18, 19].

$$P_r\left(Y = j | X = x^0\right) = \frac{1}{K} \sum_{i \in N_0} I\left(y_i = j\right) \quad (2)$$

Result and Analysis

The implementation of the SVM, LDA and KNN is done with Python Jupyter Lab, which is open source and can be used for implementation of algorithms such as classifications clustering association rules, and regression visualization.

Implementation with All Features

The machine learning algorithms such as SVM, LDA and KNN are selected to implement on UCI Parkinson data. On each implementation of these algorithms, accuracy, precision and recall are calculated. The accuracy on training data set is 0.97, 0.93 and 1 on KNN, LDA and SVM, respectively. The accuracy on testing data set is 0.90, 0.90 and 0.76 on KNN, LDA and SVM, respectively. The precision and recall on KNN are 88.0, 100.0, LDA are 89.58, 97.72 and SVM algorithms are 75.86, 100.0, respectively. In this case, SVM gives better accuracy. Tables 1, 2 and 3 show these results.

From Table 1, with all features, computed accuracy of KNN classifier on the training data is 0.97 and on test data is 0.90. Precision is 88.0; specification and recall are 100.00.

From Table 2, with all features, computed accuracy of LDA classifier on the training data is 0.93 and on test data is 0.90. Precision is 89.58, specification is 90.91, and recall is 97.73.

From Table 2, with all features, computed accuracy of LDA classifier on the training data is 0.93 and on test data is 0.90. Precision is 89.58, specification is 90.91, and recall is 97.73.

From Table 3, with all features, computed accuracy of SVM classifier on the training data is 1.0 and on test data is 0.76. Precision is 75.86, specification and recall are 100.0.

The comparison of these algorithms in terms of accuracy with all 22 features implementations is shown in Fig. 1. It shows that SVM gives best accuracy and LDA as least accuracy.

Table 2 Confusion matrix of test data using LDA algorithm with all features

True label	Predicted label	
	Negative	Positive
Negative	TN = 10	FP = 5
Positive	FN = 1	TP = 43

Table 3 Confusion matrix of test data using SVM algorithm with all features

True label	Predicted label	
	Negative	Positive
Negative	TN = 1	FP = 14
Positive	FN = 0	TP = 44

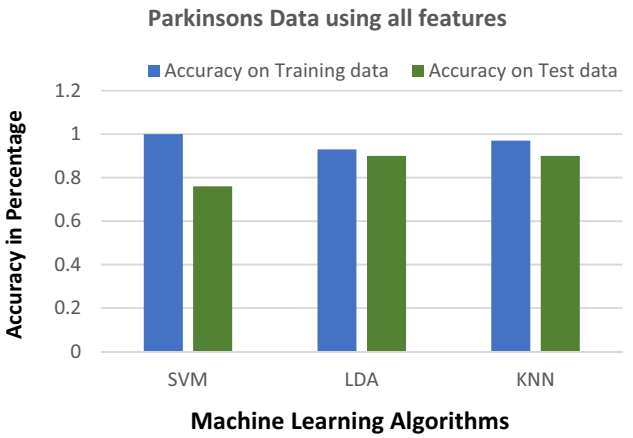

Fig. 1 Comparison of SVM, KNN and LDA Algorithm with all features

Implementation with Main 5 Features

On each implementation of these algorithms, accuracy, precision and recall are calculated. The accuracy on training data set is 0.82, 1.00 and 1.00 on KNN, LDA and SVM, respectively. The accuracy on testing data set is 0.95, 0.90 and 0.86 on KNN, LDA and SVM, respectively. The precision and recall on KNN are 0.93, 1.0, LDA are 0.91, 95, and SVM algorithms are 84.0, 100.0, respectively. Tables 4, 5 and 6 shows these results.

From Table 4, with only 5 main significant features, computed accuracy of KNN classifier on the training data is 1.00 and on test data is 0.95. Precision is 93, specificity is 100, and recall is 100.

With only 5 main significant features, accuracy of LDA classifier on the training data is 0.82 and on test data is 0.90. Precision is 91, specificity is 84, and recall is 95.

With only 5 main significant features, accuracy of SVM classifier on the training data is 1.00 and on test data is 0.86. Precision is 84, specificity is 100, and recall is 100. The comparison of these algorithms in terms of accuracy in both

Table 4 Confusion matrix of test data using KNN algorithm with only 5 features

True label	Predicted label	
	Negative	Positive
Negative	TN = 12	FP = 3
Positive	FN = 0	TP = 44

Table 5 Confusion matrix of test data using LDA algorithm with only 5 features

True label	Predicted label	
	Negative	Positive
Negative-	TN = 11	FP = 4
Positive-	FN = 2	TP = 42

Table 6 Confusion matrix of test data using SVM algorithm with only 5 features

True label	Predicted label	
	Negative	Positive
Negative	TN = 7	FP = 8
Positive	FN = 0	TP = 44

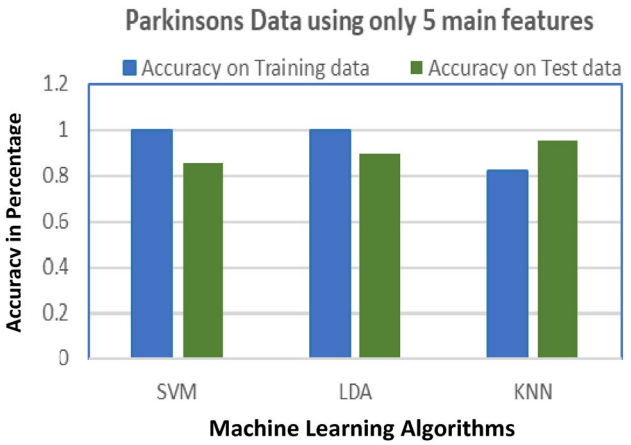

Fig. 2 Comparison of SVM, KNN and LDA Algorithm with only 5-features

implementations of 5 features is shown in Fig. 2. The accuracy of SVM and LDA is good.

Conclusion

In the current research, the effort is to experiment the machine learning algorithms to diagnose the dementia with Parkinson data. Totally, 22 subjects of the data are considered. The work initiated with 194 Parkinson's data and with the algorithms such as SVM, LDA and KNN implementation is carried out which is proof to diagnose the brain diseases through machine learning algorithms. In India or in global level, there is a scarcity of neurologists. Such diagnosis method helps to save the time of neurologists as well as patients to get the proper diagnosis in a right time. The result of implementation shows the almost 100% accuracy as all data are of Parkinson's patient data.

The implementation is with 194 data of Parkinson disease. Similarly, we can extend to diagnose the brain disease such as frontotemporal dementia and Alzheimer. These diseases not exhibit any symptoms in first stage, but, during MRI, neurologist can predict after 2nd and 3rd MRI scanning with some interval. So, such disease can be diagnosed with more accuracy using machine learning techniques. Such effort may save life of patient, and neurologist can also proceed with proper prognosis.

In future, to diagnose other dementia, subjects such as cerebral microbleeds, lacunes, cortical cerebral microfacts, white matter hyperintensities, enlarged perivascular spaces, cortical thickness, subcortical structure volumes quantified and gray matter densities have to be considered. These all values have to be collected from MRI and need to convert into numericals using appropriate software. Diagnosis of brain disease is one of the challenges to neurologist and also in the extended work; it is planned to refer these subjects to identify the disease with different machine learning algorithms referring to the large real data set. The correlation between the subjects has to be computed to investigate the association of the subjects in the progression of the disease. This work is useful in diagnosing the brain diseases with accuracy and also to distinguish these diseases as they have overlapping symptoms and attribute values.

The present paper surveys and proposes the diagnosis model for the brain disease such as Parkinson. The other diseases such as FTD disease with a combination of different feature selection algorithm and classification algorithms. The main objective of research is to get the best method to dignose the disease with more accurate and to get high performance, different combinations with different can experiment. This benefits society in the diagnosis of disease at the early stages, thus helping the clinician. With huge data set will be tried to more accurate.

References

1. WHO. Neurological disorders public health challenges; 2006.
2. Filippi M, Agosta F, Ferraro PM. Charting frontotemporal dementia: from genes to networks. J Neuroimaging. 2016;26(1):16–27.
3. http://youtube.com/watch?v=qETS3pX3Y50. Accessed 2 Sept 2018.
4. Ling H. Clinical approach to progressive supranuclear palsy. J Mov Disord. 2016;9(1):3. https://doi.org/10.14802/jmd.15060.
5. Tsanas A, Little MA, McSharry PE, Spielman J, Raming LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng. 2012;59(5):1264–71.

6. Benba A, Jilbab A, Hammouch A. Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines. Int J Electr Eng Inform. 2015;7(2):297.
7. Valli A, Jiji GW. Parkinsons disease diagnosis using image processing techniques a survey. Int J Comput Sci Appl. 2014;4(6):55–67.
8. Tan B, Venketasubramanian N, Vrooman H, Cheng C-Y, Wong TY, Chen C, Hilal S. Haemoglobin, magnetic resonance imaging markers and cognition: a subsample of population-based study. Alzheimers Res Ther. 2018;10:114.
9. Ismail R, Eltomey M, Mahdy A, Elkattan A. Hippocampal volumetric variations in the normal human brain by magnetic resonance imaging (MRI). Int J Anat Var. 2017;10(3):33–6.
10. Möller YAL Pijnenburg, van der Flier WM, Versteeg A, Tijms B, de Munck JC, Hafkemeijer A, Rombouts SA, van der Grond J, van Swieten J, Dopper E, Scheltens P, Barkhof F, Vrenken H, Wink AM. Alzheimer disease and behavioral variant frontotemporal dementia: automatic classification based on cortical atrophy for single-subject diagnosis. Radiology. 2016;279(3):838–48.
11. Lee K, Lee YM, Park J-M, Lee B-D, Moon E, Jeong H-J, Kim SY, Chung Y-I, Kim J-H. Right hippocampus atrophy is independently associated with Alzheimer's disease with psychosis. Jpn Psychogeriatr Soc. 2019;19(2):105–10.
12. Lau AY, Mok V, Lee J, Fan Y, Zeng J, Lam B, Wong A, Kwok C, Lai M, Zee B. Retinal image analytics detects white matter hyperintensities in healthy adults. Ann Clin Transl Neurol. 2018. https://doi.org/10.1002/acn3.688.
13. McCarthy J, Collins DL, Ducharme S. Morphometric MRI as a diagnostic biomarker of frontotemporal dementia: a systematic review to determine clinical applicability. Neuroimage Clin. 2018. https://doi.org/10.1016/j.nicl.2018.08.028.
14. Garrard P, Rentoumi V, Gesierich B, Miller B, Luisa M, Tempini G. Machine Learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex. 2014;55:122–9.
15. Baitharu TR, Pani SK. Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. In: International conference on computational modeling and security, CMS; 2016.
16. Marshal S. Machine learning an algorithm perspective. 2nd ed. Boca Raton: CRC Press; 2015.
17. Russell S, Norvig P. Artificial intelligence. 3rd ed. Upper Saddle River: Prentice Hall; 2010.
18. Tharwat A, Ibrahim A, Gaber T, Hassanien AE. Linear discriminant analysis: a detailed tutorial. AI Commun. 2017;30(2):169–90. https://doi.org/10.3233/AIC-170729.
19. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. Springer texts in statistics. New York: Springer; 2013. https://doi.org/10.1007/978-1-4614-7138-7.