International Journal of Engineering & Technology, *7 (4) (2018) 2795-2799*

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.17222
Research paper

Prediction of dyslexia using support vector machine in distributed environment

Jothi Prabha A [1] *, Bhargavi R [2], Ramesh Ragala [3]

[1] *Research Scholar in School of Computing Sciences and Engineering, Vellore Institute of Technology, Chennai*
[2] *Associate Professor in School of Computing Sciences and Engineering, Vellore Institute of Technology, Chennai*
[3] *Assistant Professor in School of Computing Sciences and Engineering, Vellore Institute of Technology, Chennai*
**Corresponding author E-mail: jothiprabha@gmail.com*

Abstract

Dyslexia is a learning disorder characterized by lack of reading and /or writing skills, difficulty in rapid word naming and also poor in spelling. Dyslexic individuals have great difficulty to read and interpret words or letters. Research work is carried out to classify dyslexic from non-dyslexics by various approaches such as machine learning, image processing, understanding the brain behavior through psychology, studying the differences in anatomy of brain. In addition to it several assistive tools are developed to support dyslexics. In this work, brain images are used for screening individuals who have high risk to dyslexia. This work also motivates the application of machine learning in distributed environment. The proposed predictive model uses the machine-learning algorithm Support Vector Machine (SVM). The model is designed in Apache SPARK framework to support voluminous data. The prediction accuracy of 92.5% is achieved using SVM.

Keywords: *Dyslexia; Image Processing; Support Vector Machine*

1. Introduction

1.1. Dyslexia

Dyslexia is described by a particular reading disability. Dyslexia is not a problem of visual disability, keenness or mental improvement in general. Dyslexics have abnormal brain anatomy. Also the language processing area of the brain is damaged [1]-[4]. Side effects of dyslexia are not same and can vary by language. It is nothing but an illness and untreatable. People with dyslexia need to live with it. It affects individuals of all age groups and do not depend on someone's IQ. It affects about 10% of the total population [5]. Predicting dyslexia is costly and requires a clinical or professional expert to deal with. The various methodologies used to diagnose dyslexia are machine learning, image processing in conjunction with machine learning, test-based assessments and assistive tools.

1.2. Machine Learning

Machine Learning is a method that is utilized to take in and distinguish new patterns of data from existing huge information. It allows researchers and information experts to viably recognize the strategies and plans to be conceived. To develop strategies to perform clustering, regression and classification using relevant data hidden in huge data repositories

Machine learning algorithms are broadly classified as directed and unsupervised learning. Supervised learning is the machine learning activity that infers a function from data whose class labels are already known with training examples. Unsupervised learning is a task that infers a function from data whose class labels are unknown. Machine learning is a rising trend in healthcare that helps medicinal specialists for better investigation, anticipation and treatment of people. There exist many machine learning models, every one of which performs prediction in different styles. Selecting a fitting machine learning algorithm is a one of major and complex task.

Support Vector Machine is a classifier that use kernel for pattern analysis for ranking raw data, clustering and classification of data. It creates a model that can classify a given new dataset to some class using a set of training examples. It is a non–parametric supervised learning method suitable for data which has many features.

In this paper, machine learning method SVM is implemented on brain images. MRI Functional Magnetic Resonance Imaging (fMRI) images were used for screening of dyslexia. These images are used for estimating the cerebrum activities. fMRI images are used to monitor the functioning of the mind by ensuing the adjustments in blood oxygenation. It has been observed that oxygen stream is more in control brain as opposed to a dyslexic one. Other image acquisition procedures include Positron Emission Tomography (PET), Electroencephalography and Computed Tomography (CT)

For enormous/immense information, storage and execution are significant concerns. These problems can be minimized by going for a distributed approach. Apache Spark is a quick open source framework designed for big data analytics. It is a fast open source framework suitable for handling iterative and interactive applications. It holds the properties of Map-Reduce, for example, Fault tolerance, scalability and data location. It has data flow model with Resilient distributed databases (RDD's). Apache SPARK works on Java, R, Python and Scala programming languages. It has Machine-learning library which accompanies Support Vector Machine classifier. Linear SVM is utilized as a part of this work.

Following sections emphasize on screening of dyslexia. Section 2 explains the literature work. Section 3 describes the proposed model with experimental results recorded and conclusion in section 4 and section 5 respectively.

2. Literature survey

Most of research work can be broadly categorized into 2 dimensions 1. Using Machine Learning algorithms b. image processing 2. Design tools for supporting dyslexics. Most of the research work use machine learning only for screening dyslexia. In this work, image processing and machine learning is combined for screening of dyslexia.

2.1. Screening for dyslexia using image processing and machine learning

This work proposes holistic predictive models for risk prediction. Various Data mining techniques are applied parallel on big/voluminous data to predict the risk of heart failure. Prediction model was built using Random Forest and Logistic Regression [6]. The data collected included structured and unstructured data. Classification is done parallel using Apache Hadoop framework which uses distributed computing [6].

Machine learning algorithms such as Decision Tree and Map Reduce are used on real life clinical data which has both structured and unstructured data for better prediction of various diseases. The experiment was done for prediction of chronic illness [7]. This work has concentrated on dealing with both structured and unstructured data. Compared to other existing algorithms, this work has high accuracy of 94.8% and faster convergence than that of Convolution Neural Networks (CNN) [7].

A software application called Lex for screening of phonological dyslexia is developed. Phonological dyslexic have problem in interpreting part of the word. Parents can use this tool to check whether their child is prone to risk of dyslexia or not [8]. It was found that tests of oddity and rise time were the best predictors of children with dyslexia. The software acts as a phonological marker for prediction of dyslexia.

A Novel based approach for visualizing reading difficulties was proposed in this work. Most of the tools help in prediction of dyslexia but this work focusses on understanding the aspects of the difficulties [9]. A web-based tool is designed to create descriptive visualizations for students to understand the learner model, increase awareness, support reflection, increases self-regulation to understand.

A novel technique is proposed that uses K-Means classifier, Decision Trees, Weighted Associative Classifier (WAC) for identifying the causes of heart disease. The conglomeration of K-Means, Decision Tree and WAC has shown better accuracy over existing heart disease prediction models [10]. MapReduce with Hive Database in Hadoop is used for implementation.

Prediction of dyslexia is done using positron emission tomography (PET) on 3 situations – repetition of real words, pseudo words, and rest. In both dyslexic and non-dyslexics there is no much difference in auditory processing of words and pseudo words. Dyslexic group was noticed to have less activation in right temporal lobe [11]. This deficit is due to auditory repetition and was not detected in a previous study of reading which used the same sets of stimuli.

The benefits obtained by health care from big data analytics are analyzed. Diabetics' data is used for experimentation as it has voluminous and complex data. Data driven services to patients is achieved by using MongoDB using Hadoop framework from the electronic medical records (EMR) [12].Unstructured data can be handles well using big data analytics.

Magnetoencephalography (MEG) was used to observe changes in cortical activity in dyslexic patients while reading words and sentences. It has been identified that in dyslexics, semantic analysis are sub served by the intact right hemisphere whereas non dyslexics by the left [13]. The observed variability shows the importance estimating consistency of brain activity the importance of estimating consistency of brain activity both within and between measurements in brain-damaged individuals.

A big data solution was proposed to handle healthcare problems. The major problem in healthcare is identifying the similarities between patients .Existing solutions does not support various types of data sources and they are not adaptive too .The proposed model is distributable and scalable and it solves the patient similarity problems using MapReduce framework [14]. It works on different types of data and shows high confidence and low execution time.

This model observes the motor and visual difficulties in dyslexic kids. The extent of impact on literacy difficulties is predicted. The experiments were done on kids of 2-4 years old. Traditional case study approach and logistic regression is used for building prediction model. Few predictors identified for poor learning are rapid naming problem, short term memory and phonological problems [15]. There is no single cause for dyslexia and they support multiple deficit view.

A new neurological model of dyslexia is proposed which study about the combined effect of phonological deficit and motor syndrome [16]. Cortical anomalies usually cause phonological deficit which can induce sensory pathways and also extend to other areas including cerebellum which can badly affect the reading capability. The proposed model deals with specific language impairment only and can be extended to domain specific disorders.

MapReduce software framework along with traditional data mining techniques, gives an effective way to handle data with high speed, volume and accuracy [17]. A comprehensive automated mobile healthcare system is designed for prediction of diseases. It assists the user for self-health care and diagnosis support for doctors. Only partial model was implemented due to time limit. Future scope is to build the complete model with voluminous data [17].

Neuroscience shows that sensory perception plays an important role in cognitive development. Still research is going to study the correlation between sensory perception and phonological processing which is prime factor for dyslexia [18]. It has been observed that auditory sensory impairments could lead to phonological impairment. The study was conducted on dyslexic and non-dyslexic children in three different languages English, Chinese and Spanish. Rise time acts a major predictor for reading acquisition [18]. Traditional remediation strategies such as music and basis of rhythm can be useful for linguistic and phonological development.

A healthcare smart home prediction model has been designed using simple linear regression and MapReduce framework .The model identifies the possible anomalies based on the data from wireless sensors. MapReduce was used for parallel processing of the huge dataset from wireless sensors. This model can predict early chronic diseases in elders and can be used to monitor their health over a period of time. They can predict the risk of chronic diseases before in hand by continuous monitoring [19]. Future research scope is to build learning models for clustering voluminous data obtained from sensors.

Prediction of heart disease is proposed using Hadoop Map reduce platform. Heart disease is chosen for research, as it's a major chronic disease worldwide. Improved version K-means clustering algorithm and Decision tree algorithm is used for classification. The proposed system is helpful for predicting various attributes or factors like pain in chest, cholesterol level, patient age, blood pressure levels and many more that may cause heart disorder [20]. This system can be used to assist clinical decision making.

This study shows that in addition to phonological deficits, various low level disabilities such as sensory and visual processing causes more risk for developmental dyslexia. Recent computational studies suggest that spatial attention plays a major role in reading and decoding of words [21]. Dyslexic who are incapable of word decoding also had visual and multisensory deficits. Recent study suggests temporal-parietal dysfunction as one major cause for dyslexia. The results give a new way for early prediction of dyslexia in children.

Neglect dyslexia is a reading disability acquired as a result of brain injury which is damage in selective attention. Neglect dyslexics may tend to ignore the left part of the book while reading and also beginning letters or words of a sentence [22]. The convergence of neuropsychological findings and computational modeling shows attention in visuospatial processing, and also support hybrid view for attentional selection which could be early or late. It is a new predictor which was not used in any other recent studies.

Causative theories suggest that dyslexia is a genetic disorder which could an outcome of more than one risk factors. This causes early prediction of dyslexia to be tough, which is the outcome of multiple risks. The study was conducted on preschoolers, 3 years and 8 years old children. Logistic regression was used to create prediction models [23]. Phonological awareness and language skills were major predictors in school children for screening of dyslexia.

A Prediction model for heart disease using apache Spark is proposed. The mail goal of this work is to predict the probability of heart diseases using less number of features or attributes. The data is placed Hadoop distributed file system (HDFS) and classification is done based on the features. Random forest classification is done using apache spark framework [24] .The proposed model gives a huge opportunity for health care analysts for better clinical diagnosis.

Dyslexia causes are grouped into three streams phonology, vision and multisensory. In this work phonological and deep dyslexia is analyzed. Total of 29 assessments were conducted to study the relation between these disorders. It has also been observed that many exhibit deep dyslexia without any semantic deficits. Hence it is not the only prime predictor for dyslexia [25] .Results show that phonological and semantic deficits are root cause for deep dyslexia.

3. Methodology

3.1. Architecture

The below Fig.1 describes the proposed architecture for prediction of dyslexia. The later section explains the details of implementation of the proposed model.

Working nature of Spark, ML and SVM:

Apache Spark MLlib is a distributed and scalable machine learning framework which comes with Spark. It has many machine learning algorithms and utilities to simplify large and complex machine learning tasks. MLlib supports linear SVM which is used in this work. SVM is used in this work for classification of Image dataset through feature extraction and pattern matching. Pattern analysis algorithms in general are used to study different kinds of relations such as correlations and classifications present in datasets. The structures in images are required to be transformed into an n-dimensional vector. Each coordinate is represented by a numerical value. SVM does not compute each and every coordinates of all objects. Instead, it uses a kernel, which computes the dot product of the two image vectors. From this transformed data, a decision boundary is identified between the different possible outputs. SVM is majorly used for classification and also linear regression.

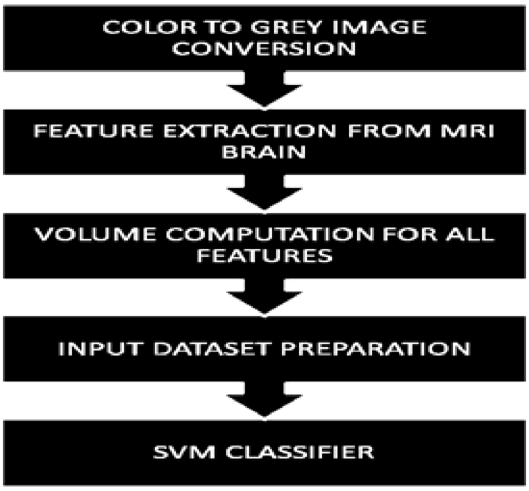
COLOR TO GREY IMAGE CONVERSION
FEATURE EXTRACTION FROM MRI BRAIN
VOLUME COMPUTATION FOR ALL FEATURES
INPUT DATASET PREPARATION
SVM CLASSIFIER

Fig. 1: Architecture.

In this work, linear SVM is used for Image classification. Linear SVM looks for finding a hyperplane decision boundary on training dataset which belong to different classes. Hyperplane is to perform linear separation of data points which belong to different classes. In Linear SVM data can be classified into two classes. The hyperplane acts as decision boundary that segregates the dataset into two parts: one part for class '0', and the other part for class '1'. [22].This applies only to dataset having two data classes. Here we use ‘0’ for non –dyslexic and ‘1’ for dyslexic.

The image dataset in this work is a set of colored images all having a fixed pixel size. The images are converted to grey scale before inputting to SVM. The image dataset has around 150 brain images (MRI).50 out of them are diagnosed as dyslexic. The aim of this work is to implement linear SVM to classify whether a given brain image is dyslexic or non-dyslexic from the brain images. These images serve as training and test data for SVM. [22].

Good feature selection and extraction play a vital role in image classification, image processing and image retrieval. Images are diverse with different categories that vary in size, shape, appearance, brightness, orientation, etc. Traditional models used Histogram of Gradients, Scale Invariant feature transform and Local binary patterns for feature extraction. Integrating these methods with Linear SVM, discriminative classifier of kernel, these hybrid models have shown better results

In this work, three kinds of features are extracted from the MRI brain image: Grey matter volume, White matter volume and cortical thickness. The grey matter volume is calculated by obtaining the RGB value of grey scale image. It eliminates the white and black region of the image and calculates the grey region volume from that. In the same way the white region of the image is calculated. Fig. 2 shows the grey matter and cortical thickness of a brain.

The fMRI brain images of dyslexics as in Fig. 3 are used in this research work. Initially the RGB values of pixels are calculated. The pixel which contains black and white intensity is removed, and the pixels that contain grey images are added to give the grey matter volume. The coloured brain images are scaled, converted into Grey scaled images. The grey scaled images are later converted to vectors. The vectors will be written to a text file. The vectors will be used by Apache Spark’s MLlib program or Linear SVM as training and test data. In this work, vectors for dyslexic and non-dyslexic brain images are created. Training sets for dyslexic and non-dyslexic are generated as .csv files. Now we have two .csv files having Class 0 (Dyslexic) and Class 1 (Non-dyslexic). We can further break down individual files for training and test data. The .csv files are given as input to SVM. Out of 50 images, there are 46 passed test cases and [4] failed test cases with an accuracy of 92.5%. Table 1 shows the confusion matrix of the results and accuracy of the model.

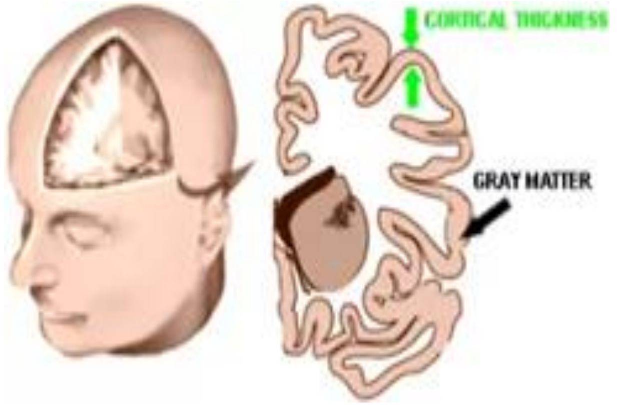

Fig. 2: Gray Matter and Cortical Thickness.

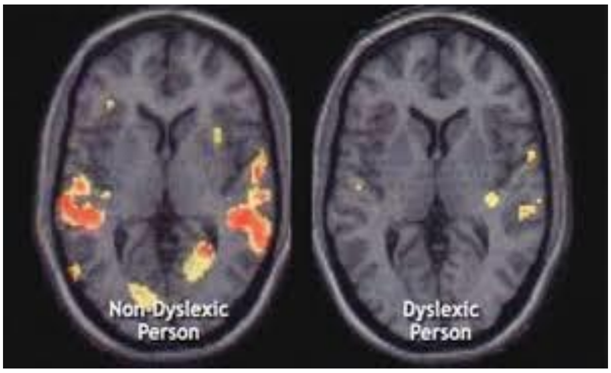

Fig. 3: Dyslexic and Control Brain While Reading.

4. Experimental results

In this research work, SVM is used for prediction of dyslexia using brain images. The features input for SVM are White matter, Grey matter and Cortical Thickness. This research work implements Apache SPARK an in-memory framework that addresses the storage processing and executions concerns of huge data.

The dataset included 150 brain MRI images of age group ranging from 24-35. 50 among them are diagnosed to have dyslexia. The adult brain images are chosen for the research, as they would have crossed the developmental reading stage as they were exposed to many kinds of study materials or methods. Validation model is stratified k-fold cross validation.

For prediction of Dyslexia, firstly images are converted to gray scale. Three features grey matter, white matter and cortical thickness are extracted from the images and are given as input vectors to the SVM. Prediction accuracy is observed to be 92.5%. It can be improved further by including more features. Table 1. and Table 2. shows the confusion matrix and experimental results respectively.

Table 1: Confusion Matrix

N=150	Predicted : 0 –Non Dyslexic	Predicted : 1 – Dyslexic	
Actual : 0 - Non – Dyslexic	TN=96	FP=4	100
Actual : 1 – Dyslexic	FN=4	TP=46	50
	100	50	

Table 2: Experimental Results

Experimental results	
Accuracy	92.5 %
Misclassification Rate	7.5 %
True Positive Rate	95%
False Positive Rate	5%
Specificity	90%
Precision	95%
Prevalence	50%

5. Conclusion

In this research paper, prediction of dyslexia is done from brain images. Linear SVM from Apache Spark MlLib is used to build the prediction model is distributed or parallel approach. The accuracy achieved is 92.5% with high specificity which is acceptable. This classification model can be useful to medical practitioners to distinguish between dyslexics and non-dyslexics. The accuracy can be improved further to by testing on a huge dataset with more number of features. Future scope of this research work is to improve the accuracy by considering other feature extraction methods and ML techniques.

References

[1] Hulme, C., & Snowling, M. J, “Reading disorders and dyslexia”, *Current opinion in pediatrics*, *28*(6),(2003),pp:731 https://doi.org/10.1097/MOP.0000000000000411.

[2] Rello, Luz, Abdullah Ali, and Jeffrey P. Bigham. "Dytective: toward a game to detect dyslexia." *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, (2015), pp. 307-308 https://doi.org/10.1145/2700648.2811351.

[3] Elnakib, Ahmed, et al. "Dyslexia diagnostics by 3-D shape analysis of the corpus callosum." *IEEE Transactions on Information Technology in Biomedicine* 16.4 (2012): 700-708. https://doi.org/10.1109/TITB.2012.2187302.

[4] Vandermosten, Maaike, Fumiko Hoeft, and Elizabeth S. Norton. "Integrating MRI brain imaging studies of pre-reading children with current theories of developmental dyslexia: A review and quantitative meta-analysis." *Current opinion in behavioral sciences* 10 (2016): 155-161. https://doi.org/10.1016/j.cobeha.2016.06.007.

[5] Kraft, Indra, et al. "Predicting early signs of dyslexia at a preliterate age by combining behavioral assessment with structural MRI." *NeuroImage* 143 (2016): 378-386. https://doi.org/10.1016/j.neuroimage.2016.09.004.

[6] Zolfaghar, Kiyana, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin, and Brian Muckian. "Big data solutions for predicting risk-of-readmission for congestive heart failure patients." In *Big Data, 2013 IEEE International Conference on*, pp. 64-71.

[7] Vinitha, S., Sweetlin, S., Vinusha, H., & Sajini, S. DISEASE PREDICTION USING MACHINE LEARNING OVER BIG DATA.

[8] Poole, Alexandra, Farhana Zulkernine, and Catherine Aylward. "Lexa: A tool for detecting dyslexia through auditory processing." *Computational Intelligence (SSCI), (2017) IEEE Symposium Series on*.

[9] Mejia, Carolina, et al. "A novel web-based approach for visualization and inspection of reading difficulties on university students." *IEEE Transactions on Learning Technologies* 10.1 (2017): 53-67. https://doi.org/10.1109/TLT.2016.2626292.

[10] Rajalakshmi, K., and K. Nirmala. "Heart disease prediction with mapreduce by using weighted association classifier and k-means." *Indian Journal of Science and Technology* 9, no. 19 (2016). https://doi.org/10.17485/ijst/2016/v9i19/93827.

[11] McCrory, E., Frith, U., Brunswick, N., & Price, C.”Abnormal functional activation during a simple word repetition task: A PET study of adult dyslexics”, *Journal of Cognitive Neuroscience*, (2000), *12*(5), 753-762. https://doi.org/10.1162/089892900562570.

[12] Basco, J. Antony, and N. C. Senthilkumar. "Real-time analysis of healthcare using big data analytics." *IOP Conference Series: Materials Science and Engineering*. Vol. 263. No. 4. (2017) IOP Publishing.

[13] Laine, Matti, et al. "Brain activation during reading in deep dyslexia: an MEG study." *Journal of Cognitive Neuroscience*12.4 (2000): 622-634. https://doi.org/10.1162/089892900562381.

[14] Barkhordari, M., & Niamanesh, M., “ScaDiPaSi: an effective scalable and distributable MapReduce-Based method to find patient similarity on huge healthcare networks”, *Big Data Research*, *2*(1), (2015), 19-27. https://doi.org/10.1016/j.bdr.2015.02.004.

[15] Carroll, J. M., Solity, J., & Shapiro, L. R., “Predicting dyslexia using prereading skills: the role of sensorimotor and cognitive abilities”, *Journal of Child Psychology and Psychiatry*, *57*(6), (2016), 750-758. https://doi.org/10.1111/jcpp.12488.

[16] Baltimore, GD Rosen, and York Press. "A neurological model of dyslexia and other domain-specific developmental disorders with an associated sensorimotor syndrome." (2006).

[17] Li, D., Park, H. W., Batbaatar, E., Piao, Y., & Ryu, K. H. ,"Design of health care system for disease detection and prediction on Hadoop using DM techniques", In *The 2016 World Congress in Computer Science, Computer Engineering, & Applied Computing (WORLDCOMP 2016), Las Vegas, USA*.

[18] Goswami, U., Wang, H. L. S., Cruz, A., Fosker, T., Mead, N., & Huss, M., "Language universal sensory deficits in developmental dyslexia: English, Spanish, and Chinese", *Journal of Cognitive Neuroscience*, *23*(2), (2011), 325-337. https://doi.org/10.1162/jocn.2010.21453.

[19] Mahmoud, S. M., & Abdulabbas, T. E, "Multiple MapReduce functions for health care monitoring in a smart environment", In *E-Health and Bioengineering Conference (EHB), 2017* (pp. 507-510).

[20] Mane, T. U, "Smart heart disease prediction system using Improved K-means and ID3 on big data.", In *Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on* (pp. 239-245).

[21] Facoetti, A., Trussardi, A. N., Ruffino, M., Lorusso, M. L., Cattaneo, C., Galli, R., & Zorzi, M. "Multisensory spatial attention deficits are predictive of phonological decoding skills in developmental dyslexia."*Journal of cognitive neuroscience*, *22*(5), (2010), 1011-1025. https://doi.org/10.1162/jocn.2009.21232.

[22] Mozer, M. C., & Behrmann, M.,"On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia." *Journal of Cognitive Neuroscience*, *2*(2), (1990), 96-123. https://doi.org/10.1162/jocn.1990.2.2.96.

[23] Thompson, P. A., Hulme, C., Nash, H. M., Gooch, D., Hayiou-Thomas, E., & Snowling, M. J.," Developmental dyslexia: predicting individual risk." *Journal of Child Psychology and Psychiatry*, *56*(9), (2015) 976-987. https://doi.org/10.1111/jcpp.12412.

[24] Saboji, R. G. ,"A scalable solution for heart disease prediction using classification mining technique.", In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing* IEEE *(ICECDS)(2017)* (pp. 1780-1785).

[25] Crisp, J., & Lambon Ralph, M. A. "Unlocking the nature of the phonological–deep dyslexia continuum: The keys to reading aloud are in phonology and semantics." Journal *of Cognitive Neuroscience*, *18*(3), (2006), 348-362. https://doi.org/10.1162/jocn.2006.18.3.348.

[26] Image Classification using Apache Spark with Linear SVM http://blogs.quovantis.com/image-classification-using-apache-spark-with-linear-svm/