An Effective Feature Selection and Classification Technique based on Ensemble Learning for Dyslexia Detection

Tabassum Gull Jan[1(corresponding author)], Sajad Mohammad Khan[2]

[1]PhD Scholar, Department of Computer Science, University of Kashmir, Srinagar (J&K)

[2]Senior Scientist, Department of Computer Science, University of Kashmir, Srinagar (J&K)

Email: [1]{tabassumgull.scholar@kashmiruniversity.net}

[2]{sajadkhan111@rediffmail.com}

Abstract: Dyslexia is the hidden learning disability where students feel difficulty in attaining skills of reading, spelling, and writing. Among different Specific Learning disabilities, Dyslexia is the most challenging and crucial one. To make dyslexia detection easier different approaches have been followed by researchers. In this research paper, we have proposed an effective feature selection and classification technique based on the Voting ensemble approach. Our proposed model attained an accuracy of about 90%. Further Comparative analysis between results of various classifiers shows that random forest classifier is more accurate in its prediction. Also using bagging, the Stacking approach of ensemble learning accuracy of classification was further improved.

Keywords: *Ensemble Learning, Feature Selection, Voting, Stacking, Bagging, Dyslexia*

1. Introduction

Dyslexia is a neurological disorder, manifesting as difficulty in reading, comprehending, writing, spelling, using other language skills, and calculations. This disability results from differences in the way a person's brain is developed. Children with such learning disabilities are often more intelligent. If such students are taught using traditional classroom instructions and rules, they may have difficulty in reading, writing, spelling, reasoning, recalling, and/or organizing information. Among different types of learning disabilities, dyslexia is the most common. Dyslexia is a hidden learning disability where students usually experience difficulties with other language skills, such as spelling, writing, and pronouncing words. The intensity of difficulty a child with dyslexia is having varies due to inherited differ-

ences in brain development, as well as the type of teaching the person receives. Their brain is normal, often we say very "intelligent," but with strengths and capabilities in areas other than the language area.

Between 5 - 15% of people in the United States i.e. about 14.5 to 43.5 million children and adults in America have dyslexia as per the reports of the Society for Neuroscience on "Dyslexia: What Brain Research Reveals About Reading". In the last few years, the number of children who are labeled as learning disabled (LD) or dyslexic in India has increased exponentially [1], and currently about 5-15% of the school-going children's have a learning disability and from these learning-disabled students around 80% are dyslexic. According to the "Dyslexia Association of India," 10-15% of school-going children in India suffer from some type of Dyslexia. The incidence of dyslexia in Indian primary school children has been reported to be 2-18% [2][3]. The prevalence of SLD in school children in south India is 6.6% [4].

With advancements in the field of technology and artificial intelligence, the researchers have tried very hard to design various techniques and for distinguishing dyslexic people from non-dyslexic ones. These include designing various machine learning approaches, application of various image processing techniques, designing various assessment and assistive tools to support and ease the problems encountered by dyslexic people . Since machine learning techniques are broadly used in dyslexia detection. The scope of this paper is to improve the accuracy of the existing techniques by the use of more effective feature selection and classification techniques. In this paper, we have proposed an ensemble approach for effective feature selection and classification of dyslexic students from non-dyslexic ones. The paper is summed up in the following sections: Section 1 and 2 covers the introduction and related works. Section 3 briefly explains the Dataset and the proposed technique. Subsequently, the results of the technique are explained in detail and results are penned down in Section 4. Lastly, the summary of the work is concluded with acknowledgments and references.

2. Literature Review

Machine learning has been widely used in medical diagnosis nowadays. With the speedy increase in the era of artificial intelligence, deep neural networks have been used in variety of applications for accurate classification with an automated feature extraction [16]. Hybrid models developed using machine learning have been used to predict emotions involved in the child’s behavior [17]. Another domain is ELM (Extreme Learning Machine) which has become new trend in learning algorithms nowadays. In ELM, to have good recognition rate in less amount of

time researchers are trying hard to attain stable versions of biases for classification tasks [18]. Different wavelet transform Techniques have been used in feature detection of images and signals [19]. Nowadays EIT (Electrical Impedance Tomography) of medical images plays an important role in medical application field [20]

. Dyslexia has been treated by researchers differently. Different eye-tracking measures employed with machine learning approaches were designed and the aim was to screen the readers with and without dyslexia based on input taken from interaction measures when readers were supposed to answer questions testing their level of dyslexia [6][9]. In 2018, Khan et al proposed a diagnostic and classification system for kids with dyslexia using machine learning. The overall accuracy of system was 99% [8]. Further, Student Engagement Prediction Using machine learning was done and an accuracy of 97.8% was obtained [7]. Dyslexia detection has been done using machine learning have been done various approaches like eye-tracking measures, EEG based, image processing based, MRI based, fMRI based, questionnaire based, game-based, an assistive tool based [10-13]. Further dyslexia detection using handwritten images was studied and an accuracy of 77.6% was achieved [13].Game-based approach was used by researchers to detect dyslexia [5][14-15]. Here readers were asked to answer different questions set accordingly to test different cognitive skills that cover working memory, Language Skills, and Perceptual Processes.

3. Proposed Method

The proposed method is an ensemble-based approach for dyslexia detection in which features are selected via three techniques. The three feature selection techniques used are Select K Best features, Mutual Information Gain, and Recursive Feature Elimination. The features selected via all approaches are then subjected to Classification using five machine learning models. The five machine learning algorithms used are Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, and K Nearest Neighbor Classifier. The results of the best classifier from the selection techniques are taken and supplied as input to the voting classifier. The results are aggregated by maximum voting to yield the final result as shown in Table 2. Furthermore, we have implemented stacking on these algorithms which were aimed to increase the accuracy of classification. Comparisons with existing work were shown by applying the bagging approach on three different datasets obtained via three feature selection approaches with the same configuration.

The proposed technique is the design for an efficient classification technique for dyslexia. The overall framework for the proposed system is shown in Fig. 1. The

first phase of the proposed technique is to compute features from dataset for classification via three feature selection techniques namely Select K Best features based on chi squared statistical parameter, mutual information Gain and Recursive Feature Elimination method. The second phase is to train the classifiers on the basis of three set of features selected for inducing ensemble classification. The elaborated view of dyslexia classification module is shown in Fig 2. The goal of the ensemble classification is to increase the predictive accuracy.

4. Dataset

The dataset used in the research has been taken from Kaggle and is freely accessible at https://doi.org/10.34740/kaggle/dsv/1617514 . The dataset was prepared in the research work entitled "Predicting the risk of Dyslexia using online gamified test" [5]. The dataset was collected from an online gamified test based on machine learning. Dataset has two parts one meant for training the classifiers and other meant for testing purposes. About 3000 plus participants participated in collecting data for training and 1300 participants participated in collecting data for testing. Training dataset consists of 3644 samples (392 were dyslexic and 3252 were non dyslexic) and 1365 samples (148 are dyslexic and 1248 were non dyslexic) in testing. The dataset has 197 features in total. All the participants were equal or more than 12 years old.

5. Data Preprocessing

The raw data taken from the dataset is transformed into the classifier understandable format by the process known as data preprocessing. It is a very important and crucial step to check the quality of data in terms of accuracy, completeness, consistency, believability, and interpretability before applying it to the machine learning model. Major tasks involved are data cleaning, data reduction, data transformation. In data cleaning, incorrect, incomplete, and inaccurate data is removed. Missing values or null values are handled by either mean or by applying regression analysis. Further categorical data is converted to numerical data using Label Encoder because machine learning models use mathematical equations and take numerical data as input. The label Dyslexia of the dataset is handled using dummy encoding where we replace category Yes by 1 and No by 0 dummy variables.

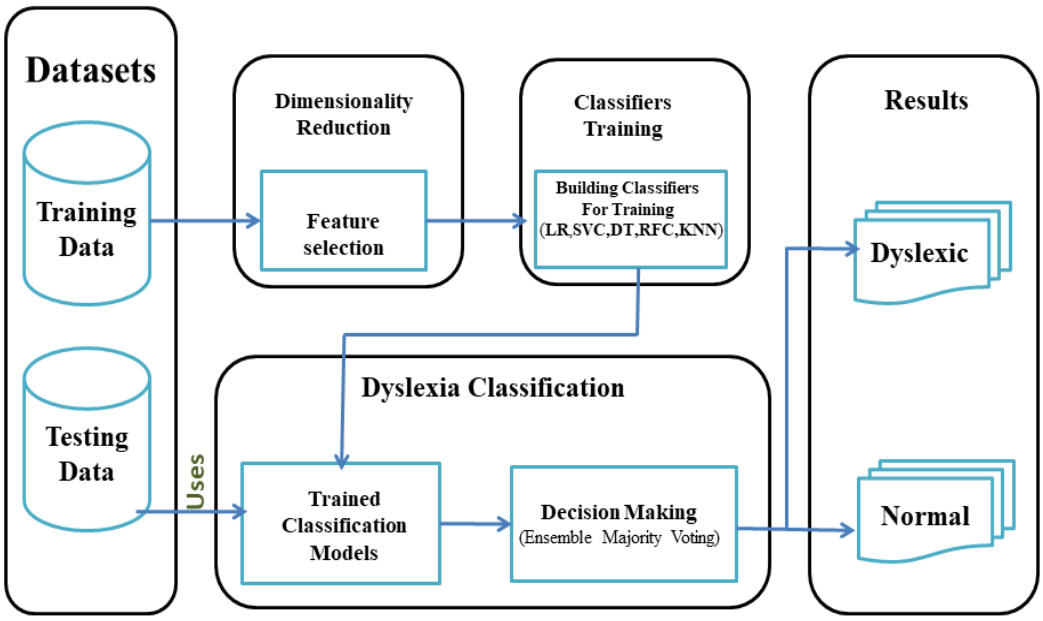

Fig. 1: The Overall framework of Proposed Technique

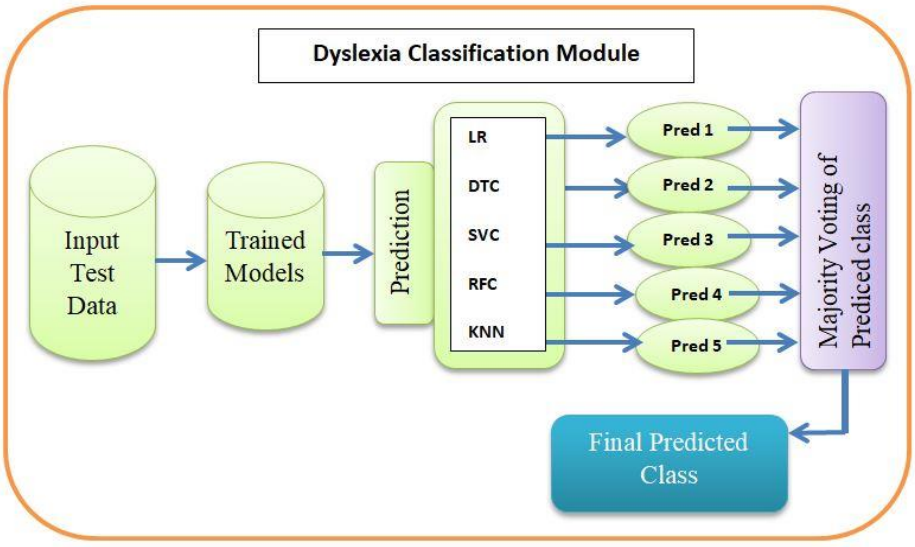

Fig. 2: Detailed Structure of Dyslexia Classification Module

6. Features Selection Techniques

Machine learning has been used for feature selection in this research work. Feature selection removes the irrelevant and less important features from the dataset that can negatively impact and reduce the overall performance of the model. Feature selection reduces the training time, dimensionality, and over fitting of the model. The features in the dataset correspond to a total of 197 features. Since the dimensionality of the dataset is large that in turn corresponds to more computational cost.

Feature Selection Techniques employed in this research work are:

I. **Mutual Information Gain –** It is defined as the "amount of information present in the feature for identifying the target value" and calculates a decrease in the "entropy values". Information gain (IG) for each attribute is calculated by considering the target values for feature selection.

II. **Select K Best Features –** In this technique K- Best Features are selected based on the highest Chi-Square Test Statistic Values. Chi-square test is a technique used to determine the relationship between the categorical variables. The "chi-square value" is calculated between each feature and the target variable, and the required number of features with the highest chi-square value is selected.

$$(X)^2 = \sum (observed\ value - expected\ value)^2\ /Expected\ Value$$

III. **Recursive Feature Elimination** (**RFE**)is a "recursive greedy optimization approach, where features are selected by recursively taking a smaller and smaller subset of features". Then, an estimating classifier is trained with each set of features, and the importance of each feature can be determined.

The above three feature Selection methods were applied to reduce the dimensionality of the dataset as shown in Table 1 (where f1 denotes feature no 1, f2 denotes feature no 2 and so on).

Table 1: Summary of Feature Selection Techniques Used

S.No	Feature Selection Method	Features extracted
1	Select K Best (K=10)	f28,f34,f118,f125,f130,f143,f148,f154,f160,f167
2	Mutual Information Gain	f3,f5,f8,f11,f13,f16,f17,f18,f20,f23,f24,f28,f32,f34,f35,f36,f41,f47, f51,f54,f57,f62,f63,f66,f67,f70,f73,f76,f77,f84,f85,f95,f96,f97,f100 ,f101,f103,f105,f110,f111,f112,f114,f118,f119,f120,f121,f123,f124, f132,f133,f134,f135,f136,f137,f138,f141,f142,f143,f144,f145,f147, f148,f149,f150,f151,f153,f154,f155,f156,f157,f158,f159,f160,f161 ,f162,f164,f165,f166,f167,f168,f169,f171,f172,f173,f174,f178,f179, f185,f186,f187, f188,f189, f190,f191, f192, f193, f196,f197
3	RFE (Recurssive Feature Elimination (K=10)	f5, f6,f18, f24, f63, f66, f144, f150, f168, f190

7. Role of Ensemble Majority Voting

Voting Ensemble classifier (often called Ensemble Majority Voting Classifier) is an ensemble machine learning Classifier that is used to combine predictions made by multiple classifier models as shown in Fig 2. This majority Voting was applied to the predictions made by the individual classifiers that were trained individually. Voting Classifier combined the predicted class outputs of all classifiers and produces the final predicted class output using Mode statistic. Hence using an ensemble voting classifier performance of the model was improved compared to single contributing classifiers.

8. Experimental Setup

A series of experiments were performed for the detection and classification of dyslexic and non-dyslexic individuals with a dataset representation procedure that allows the machine learning classifiers to learn the model quickly and more accu-

rately. The proposed method was implemented using Python programming language using machine learning libraries (Scikit-learn) on Intel Core i5 Processor having 8 GB RAM installed.

9. Results and Discussions

In this experimental study, all the experiments were carried out on the dataset [5] that contain 3644 training samples and 1365 testing samples. Also, dataset has a total of 197 columns which makes it computationally more costly to work with. Therefore, we have implemented efficient feature selection techniques to reduce the dimensionality of the dataset as well as the computational cost. We have selected the informative features via three different feature selection methods viz Select K Best features based on chi-squared statistical parameter, mutual information Gain, and Recursive Feature Elimination method. From Experimental results in Table 3-5, it is quite evident that the K Best Feature selection technique is efficient. The proposed technique via the voting ensemble approach attained an accuracy of 90.2% as shown in Table 2. In the existing approach [5] random forest classifier with 10 fold cross-validation and the number of trees= 200 was implemented on Weka 3.8.3 framework and an accuracy of 89% was attained. The experimental results of the proposed technique showed an increase in accuracy while implementing the bagging approach of ensemble learning with number of tress=200 and 10 fold cross validation as shown in Table 6. Further, the Stacking approach was implemented with (Logistic Regression, Decision Tree Classifier, Random Forest Classifier, KNN, and SVM) as base learners and Decision Tree as Meta Learner. All the three feature selection approaches were implemented, and the results of stacking further increased the accuracy and it was **90.6%**, **90.7%,** and **89.4%** shown in Table 3, Table 4 and Table 5 respectively.

Table 2: Experimental results of Proposed Technique

S. No	Classifier	Accuracy	Standard Deviation	Voting
1	Logistic Regression	0.873	0.030	
2	KNN	0.889	0.038	
3	Decision Tree	0.895	0.031	**0.902**
4	SVM	0.899	0.035	
5	Random Forest	**0.900**	0.033	

KNN(K Nearest Neighbor); SVM(Support Vector Machine)

Stacking ensemble approach was also performed and it was concluded from results that overall accuracy was further improved than voting ensemble Classifier.

Table 3: Feature Selection Method (Mutual Information Gain)

S. No	Classifier	Accuracy	Standard Deviation	Stacking
1	Logistic Regression	0.897	0.012	
2	KNN	0.889	0.008	
3	Decision Tree	0.851	0.020	**0.906**
4	SVM	0.892	0.001	
5	Random Forest	**0.898**	0.004	

Table 4: Feature Selection Method (K Best Features)

S. No	Classifier	Accuracy	Standard Deviation	Stacking
1	Logistic Regression	0.897	0.012	
2	KNN	0.889	0.008	
3	Decision Tree	0.849	0.019	**0.907**
4	SVM	0.892	0.001	
5	Random Forest	**0.899**	0.005	

Table 5: Feature Selection Method (Recursive Feature Elimination)

S. No	Classifier	Accuracy	Standard Deviation	Stacking
1	Logistic Regression	0.893	0.009	
2	KNN	0.886	0.012	
3	Decision Tree	0.834	0.028	**0.894**
4	SVM	**0.894**	0.003	
5	Random Forest	0.886	0.013	

Table 6: Results of Bagging Ensemble Approach (with number of tress=200 for Decision Tree Classifier)

S. No	Feature Selection Method	Bagging	Accuracy of Existing Approach
1	K Best Features	90.5%	89 % (All features were used)
2	Mutual Information Gain	89.5%	
3	Recursive Feature Elimination	89.2%	

From Table 3, Table 4, Table 5 and Table 6 It can be concluded that applying ensemble learning to the existing approaches showed an increase in the accuracy. Existing approach have an accuracy of 89% while proposed technique achieved improvement in accuracy (90.2% for Voting Ensemble, 90.5% for Stacking and 90.5% for Bagging)

10. Conclusion

The proposed approach presented in this paper showed accuracy of the existing approach [5] has increased. From experimental results we conclude K- Best Feature Selection approach leads to increase in the overall of accuracy of Stacking approach showed an increase as compared to other two feature section methods. By applying voting ensemble learning accuracy of 90.2% was attained. Lastly we

summarized that proposed approach out performed existing approach based on all the ensemble techniques employed, whether it be Bagging, Stacking or Voting.

11. Acknowledgments

This work is financially supported by Department of Science & Technology, Government of India under **DST INSPIRE Fellowship** Scheme bearing registration Number **IF190563.**The grant is received by Tabassum Gull Jan.

12 References

[1] S. Karande, R. Sholapurwala and M. Kulkarni, "Managing specific learning disability in schools in India", *Indian Pediatrics*, vol. 48, no. 7, pp. 515-520, 2011.

[2] M. SK, Z. I, P. N, D. S, R. B and B. SK, "Communication disabilities: emerging problems of childhood", *Indian pediatrics* ,vol. *14* ,no.10,pp. 811-815, 1977.

[3] S. Singh, V. Sawani, M. Deokate, S. Panchal, A. Subramanyam, H. Shah and R. Kamath, "Specific learning disability: a 5 year study from India", *International Journal of Contemporary Pediatrics*, vol. 4, no. 3, p. 863, 2017.

[4] S. Bandla, G. Mandadi and A. Bhogaraju, "Specific Learning Disabilities and Psychiatric Comorbidities in School Children in South India", *Indian Journal of Psychological Medicine*, vol. 39, no. 1, pp. 76-82, 2017.

[5] L. Rello, R. Baeza-Yates, A. Ali, J. Bigham and M. Serra, "Predicting risk of dyslexia with an online gamified test", *PLOS ONE*, vol. 15, no. 12, p. e0241687, 2020.

[6] L. Rello and M. Ballesteros, "Detecting readers with dyslexia using machine learning with eye tracking measures in Proceedings of the 12th International Web for All Conference", *ACM*, pp.1-8, 2015 [Online]. Available: https://dl.acm.org/doi/10.1145/2745555.2746644.

[7] S. S. A. Hamid, N. Admodisastro, N. Manshor, A. Kamaruddin, and A. A. A. Ghani, "Dyslexia adaptive learning model: Student engagement prediction using machine learning approach," in *Recent Advances on Soft Computing and Data Mining: Advances in Intelligent Systems and Computing*, vol. 700, R. Ghazali, M. Deris, N. Nawi, and J. Abawajy, Eds. Cham, Switzerland: Springer, pp. 372-384, 2018, doi:10.1007/978-3-319-72550-5_36

[8] R. U. Khan, J. L. A. Cheng, and O. Y. Bee, ``Machine learning and Dyslexia: Diagnostic and classification system (DCS) for kids with learning disabilities," *Int. J. Eng. Technol.*, vol. 7, no. 3, pp. 97_100, 2018.

[9] M. N. Benfatto, G. Seimyr, J. Ygge, T. Pansell, A. Rydberg, and C. Jacobson, "Screening for dyslexia using eye tracking during reading", *PLoS ONE*, vol. 11, no. 12, 2016, Art. no. e0165508, doi: 10.1371/journal.pone.0165508.

[10] H. Perera, M. Shiratuddin, K. Wong and K. Fullarton, "EEG Signal Analysis of Writing and Typing between Adults with Dyslexia and Normal Controls", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 1, p. 62, 2018.

[11] F.J. Martinez-Murcia, A. Ortiz, R., Morales-Ortega, P. J. Lopez, , J. L. Luque, Castillo-Barnes, J. M. Górriz "Periodogram connectivity of EEG signals for the detection of dyslexia." In Int.Work-Conf. on the Interplay Between Natural and Artificial Computation, Springer, Cham, pp. 350-359, 2019.

[12] A Jothi Prabha , R. Bhargavi , R. Ragala "Predictive Model for Dyslexia from Eye Fixation Events", International Journal of Engineering and Advanced Technology, vol. 9, no. 13, pp. 235-240, 2019.
[13] K. Spoon, D. Crandall, and K. Siek, ``Towards detecting Dyslexia in children's handwriting using neural networks," in Proc. Int. Conf. Mach. Learn. AI Social Good Workshop, 2019, pp. 1-5.
[14] L. Rello, K. Williams, A. Ali, N. Cushen White, and J. P. Bigham, "Dytective: Towards detecting dyslexia across languages using an online game", In Proc. W4A'16, Montreal, Canada, 2016. ACM Press.
[15] L. Rello, M. Ballesteros, A. Ali, M. Serra, D. Alarcon, and J. P. Bigham," Dytective: Diagnosing risk of dyslexia with a game", In Proc. Pervasive Health'16, Cancun, Mexico, 2016.
[16] Bashar, Abul. "Survey on evolving deep learning neural network architectures." Journal of Artificial Intelligence 1, no. 02 (2019): 73-82.
[17] Kumar, T. Senthil. "Construction of Hybrid Deep Learning Model for Predicting Children Behavior based on their Emotional Reaction." Journal of Information Technology 3, no. 01 (2021): 29-43.
[18] Mugunthan, S. R., and T. Vijayakumar. "Design of Improved Version of Sigmoidal Function with Biases for Classification Task in ELM Domain." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 70-82.
[19] Manoharan, Samuel. "Study on Hermitian graph wavelets in feature detection." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 24-32
[20] Adam, Edriss Eisa Babikir. "Survey on Medical Imaging of Electrical Impedance Tomography (EIT) by Variable Current Pattern Methods."Journal of ISMAC 3, no. 02 (2021): 82-95.