CLASSIFICATION AND SUBGROUP ANALYSIS OF DYSLEXIC INDIVIDUALS

AMMARA MINHAS

01-249192-023

A thesis submitted in fulfilment of the
requirements for the award of degree of
Masters of Science (Data Science)

Department of Computer Science

BAHRIA UNIVERSITY ISLAMABAD

SEPTEMBER 2021

Thesis Completion Certificate

Scholar Name: Ammara Minhas

Registration Number: 63103

Enrollment: 01-249192-023

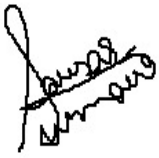
Program of Study: MS (DATA SCIENCE)

Thesis Title: CLASSIFICATION AND SUBGROUP ANALYSIS OF DYSLEXIC INDIVIDUALS

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index **17%**. that is within the permissible limit set by the HEC for the MS/M.Phil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/M.Phil thesis.

Principal Supervisor Name: Dr. Sumaira Kausar

Principal Supervisor Signature:



Date: 10 Sep 2021

Author's Declaration

I, Ammara Minhas hereby state that my MS/M.Phil thesis titled is my own work and has not been submitted previously by me for taking any degree from Bahria university or anywhere else in the country/world. At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS/M.Phil degree.

Name of Scholar: Ammara Minhas

Date: 10 Sep 2021

Plagiarism Undertaking

I, solemnly declare that research work presented in the thesis titled **" CLASSIFICATION AND SUBGROUP ANALYSIS OF DYSLEXIC INDIVIDUALS"** is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me. I understand the zero tolerance policy of the HEC and Bahria University towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS/M.Phil degree, the university reserves the right to withdraw / revoke my MS/M.Phil degree and that HEC and the University has the right to publish my name on the HEC / University website on which names of scholars are placed who submitted plagiarized thesis.

Name of Scholar: Ammara Minhas

Date: 10 Sep 2021

Dedication

I dedicate my thesis work to my supervisor,my respected teachers and my parents.

Acknowledgements

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Sumaira Kausar, for encouragement, guidance, critics and friendship. I would also like to thank Hafiz Fahad Siddique for the continuous guidance and support. Without their continued support and interest, this thesis would not have been the same as presented here.

Librarians at Bahria University also deserve special thanks for their assistance in supplying the relevant literature. My fellow postgraduate students should also be recognized for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family members.

Abstract

Analysis of dyslexia is an interesting research area in medical field through MRI scans and eye ball tracking etc and also it is very interesting research in machine learning through different models. Dyslexia is a learning difficulty that affects the ability of reading, writing, listening and learning in children and adults. It is mostly misclassified as disorder but it is basically a difficulty. Many research's in detecting dyslexic or non-dyslexic patients in many different ways were carried out by researchers, however the focus of our current study is to classify the dyslexia and further do the subgroup analysis of dyslexic patients using machine learning algorithms. Machine learning is a technique used in computer science in which models automatically learns the way humans learn. They learn from the data given to them. Machine learning algorithms are used mostly for prediction. In the current scope of research, for this purpose Different assessment techniques are used to assess students based on skills like reading, writing and sounds. The score of these assessments are further used for classification and analysis of subgroups in dyslexic patients. Supervised and un-supervised both techniques are used for classification and subgroup analysis. For classification purpose both binary class and multiclass classification is done. For binary classification purpose neural network, for multiclass classification svm and for clustering subgroup analysis k-means algorithm is used. In binary classification dyslexic and non- dyslexic patients have been classified. In multiclass classification different classes are classified and after this clustering is done using k-means algorithm to make different groups of different dyslexic level.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

CHAPTER 1

INTRODUCTION

Dyslexia is originated from a Greek word "dys" which means difficulty and "lexia" which means language. Dyslexia is a learning difficulty that affects the ability of reading, writing, listening, and learning in children and adults. It is mostly misclassified as disorder although it is a difficulty. Following are different types of dyslexia.

1. Phonological dyslexia are related to the difficulty in understanding of words by their sounds.

2. Visual dyslexia are related to the difficulty in reading and to the vision problems

3. Primary dyslexia are type of difficulties transferred genetically by either father or mother

4. Secondary dyslexia occurs when brain is not growing properly

5. Surface dyslexia means to delay in processing language, child with this take more time to decode words as they sound in English.

Patients suffering from dyslexia are slow learners. They learn and grasp things slowly though they are very intelligent aesthetically. Detection of dyslexia is considerably difficult because of unsupportive attitude of patient's parents especially with children.

Although there exists couple of problems and issues that are leading towards the increase or acting as a barrier to the appropriate treatment of dyslexic patient but some of the major reasons include unsupportive parents, education system, environment as well as less opportunities for social awareness. There are millions [1] of children globally suffering from dyslexia. Early detection and Treatment of dyslexia is very important as it helps through the betterment of the patient especially in case of children, it should be considered as a primary social responsibility and necessity as it may affect schooling and future of the child.

Figure 1.1: Understanding Dyslexia

[2]

According to the stats 12 million children in Pakistan are suffering from dyslexia not knowing about it. According to the study 15 percent to 20 percent of students in each section are suffering from dyslexia[1]. They face difficulty in each subject and hence named as dull students by the teachers. Our education is so weak that in higher classes most of the male students leave the school because of failure but no one knows the root cause. Primarily the root cause behind the failure in secondary school is weak primary school. One will be unable to understand any advance level concepts when he/she doesn't have

the basics of subject. Mostly children face difficulties in speaking, as they can't pronounce the words due to carelessness of teachers in the early stage.

In schools, students bully others based on color complexion, marks, body type and much more. The concept of favoritism is also present in our schools where usually teachers give favors to some students while ignore others and this leads to never-ending clash between the groups.

The fight against dyslexia is in reality against our traditional education and parental system. Parents and teachers should to motivate their child and student. Both must work hard for the great future of our students and our nation. Teachers should encourage students who face difficulties in study instead of shouting on them and saying cruel words and disrespecting them they must stand by them and assure them that they are with them always. Every child has its own personal skill in which there are more chances of success because every individual is not capable of doing same things and doesn't have same mind and thinking capability.

Furthermore, studies shows that the brain of dyslexic people are different compared to non-dyslexic. Dyslexic brains are organized differently due to which they function differently. Human brains are made up of a matter in the form of fiber which are responsible and helps in reading and writing, however the person with dyslexia has a different distribution of metabolic activation than normal person so they find difficulties in reading, writing, and learning. In addition to that in the dyslexic brain, there is more activity in the frontal lobe and less activity in the parietal and occipital areas of the brain. Through different functional magnetic resonance imaging it is evident that in dyslexic children there is some functional disruption in neural systems.

Advances in computer technology as well as related software application have greatly impacted the educational field. Recently, there have been a number of learning tool that could play an important role to motivate student's interest hence facilitating their performance in classrooms. Thus, a computer-

based screening is the most crucial effort as an alternative approach to the manual screening whereby dyslexic children can attend classes with special learning method suggested by a psychologist.

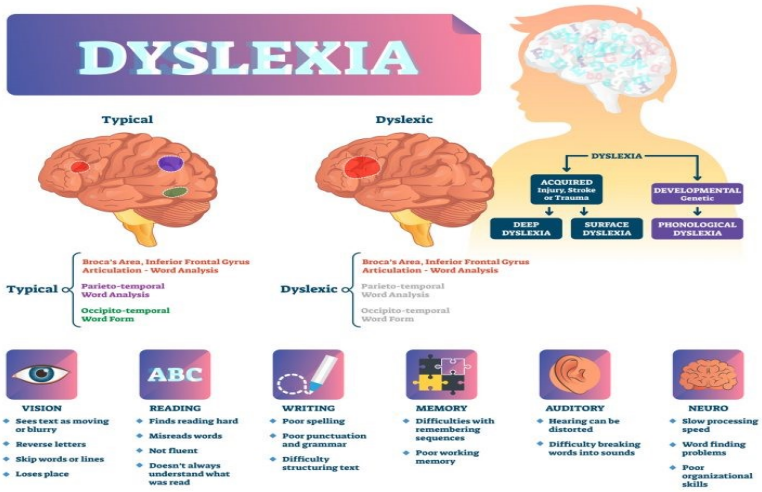


Figure 1.2: Typical and Dyslexic brain difference

[2]

Brain is a very complex and sensitive part of human body that performs different functions. The main function of the brain is to control overall body. The brain is divided into right hemisphere and left hemisphere fig:(1.5). Mostly areas which are responsible for processing of language, speech and reading are present on the left hemisphere. Mainly the Brain is made up of gray matter and white matter. The gray matter is made up of nerve cells and is responsible of processing information. Whereas white matter is present in the deeper part of brain and made up of connected fibers in myelin and is responsible for transfer of information around the brain [14].

Shaywitz et al 2004., [14] examine that dyslexic people have less gray matter in their left parietotemporal area (Part A in Fig.1.3). Less gray matter in this area means that the person is having sound related difficulties. This is area is involved in analysis of words, decoding them and mapping words and letters with their sounds. Shaywitz also examined that the left occipi-

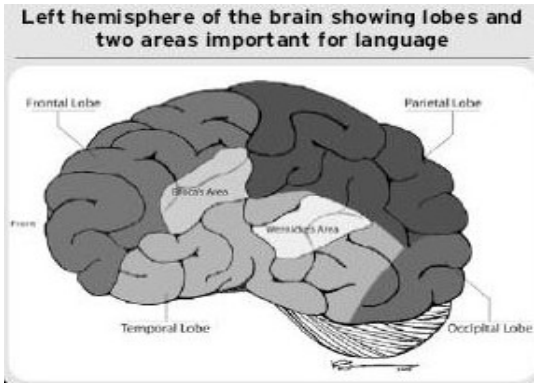


Figure 1.3: Important Areas for Language

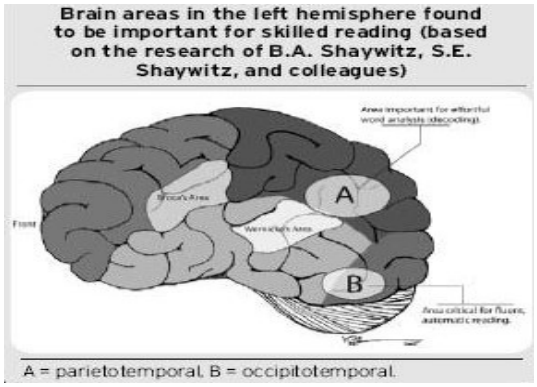


Figure 1.4: Left Parietotemporal and Occipitotemporal area

Figure 1.5: Left Hemisphere of Brain

[3]

totemporal area is important for reading. This involves reading fluently and is sensitive. fig: 1.4

1.1 Challenges in detection of dyslexia

As our educational system requires competitive reading and writing, school failure is linked to dyslexia, even if dyslexia are unrelated to overall intellect. The early diagnosis of dyslexia is critical. When dyslexia are identified, it can be treated to avoid the negative effects of the condition, such as high rates of academic failure. At the same time, diagnosing dyslexia is not

a simple process; it is costly because it requires lots of tests of different types and usually need the assistance of a professional. In addition, the symptoms of dyslexia differ depending on the language. Dyslexia is known as a "hidden handicap" since it is difficult to diagnose in languages.

1.2 Problem Statement

Dyslexia is a type of learning disorder which requires early and detailed detection. Diagnosing dyslexia must be done based on some of the most important factors like reading, writing, spelling and pronunciation difficulties which are likely to be present in that particular age group of dyslexic child. Normally detection is done manually or sometimes using automated systems, but they do not focus on different factors rather they diagnose dyslexia with only one or two factors which does not give accurate results.

Basic learning tests are required or through many aspects dyslexia can be diagnosed but unfortunately due to unawareness of this learning difficulty there are not enough facilities and automated systems to properly detect dyslexia. This cannot be determine based on any single attribute like basics of learning as well as all the associated problems and severity of similar types of dyslexia cannot also be the same. So, considering this into account the treatment, rehabilitation, and intervention of different type of dyslexia cannot be the same. Mostly people have only detected dyslexia but not explored types of dyslexia, focusing on types of dyslexia is important for better treatment.

Dyslexia testing and screening are both available and necessary. This difficulty may cause frustration, academic failure, and poor self-esteem if it is not diagnosed and treated properly. So in this proposed system identification of dyslexia is done on the basis of considering multiple important factors. This will help to generate more accurate and effective results .

1.3 Research Objective

Objective of this research is to detect dyslexics amongst non-dyslexic through supervised learning. The main objective of this research is to identify different underlying hidden patterns or similarities. Identification of these patterns are very important for further analysis. Identification of different classes is very important for dyslexic students. As each student is different from every other and level of severity is also different in learners. So for the purpose of individual treat- ment and rehabilitation, identifying different classes is very necessary as every learner needs different type of treatment.

1.4 Motivation on analysis of Dyslexia

Primary interest of us is to work on the research topic that is related to some community issue, any disease, disabilities, and/or health related domains, so after lot of research and studies detection of dyslexia was finalized. There were difficulties because not much data available on the internet in this regard, however by the passage of time we got to know that there is an organization with the name bazeecha trust working on the same domain, so I took an appointment from them and visited Bazeecha trust. After learning through couple of case studies and the social issues attached with this disease we started getting highly motivated and started this research.

1.5 Key Contributions

The key contributions of the present study are listed in the following.

- Investigation of pre-processing techniques and applying some methods for pre-processing.
- Identify best performance algorithm for binary and multiclass classification.

- Applying binary classification techniques.
- Multiclass classification for the purpose of individual rehabilitation.
- Clustering to identify underlying patterns for detailed analysis.

1.6 Thesis Organization

This document is organized as follows. Chapter 2 presents the methods used for identification and detection of dyslexia. Chapter 3 describes the method that we have adopted in order to achieve the objectives along with the key concepts behind the approaches. Chapter 4 outlines the metrics used to test our methods, describes the experiments, presents the findings we obtained and their interpretation. Chapter 5 incorporates the concluding remarks and recommendations for future work.

CHAPTER 2

LITERATURE REVIEW

There are different research dimensions in dyslexia. Many researchers have worked in these like some of them used machine learning techniques, some of them used image processing techniques using brain MRI, EEG scans and eye ball tracking systems.

Dyslexic individuals are identified based on their poor marks in different assessments. However, because symptoms differ between people, these procedures are generally time-consuming and useless for a wide group of people. As a result, researchers adopt machine learning methodologies, which are less time intensive and often inexpensive. In this scenario, a variety of tests, including reading, writing, and typing, handwriting, and a web-based game, are used to collect various types of data, including text, eye movement, MRI scans, EEG scans, and image analysis. The participants' ages ranged from 7 to 62, and their native languages were Spanish, Hebrew, Swedish, Mandarin, French, German, Polish, Malay, English, Greek, and Dutch.

2.1 Machine Learning techniques in dyslexia

In 2019, **Iza Sazanita Isa et al.,** [4] created a system based on handwritten graphics that used a pattern recognition approach. On the basis of their handwritten photographs, the pattern recognition approach indicates and extracts the features of written characters using Optical Character Recogni-

tion (OCR). The accuracy of the classification was achieved by comparing the automatic accurate identification and manual measurement of 73.33 percent.

Iwabuchi et al in 2017., [5] aims to provide students with a simple way to check issues related to their reading and writing. In this study, they have used evaluation method URAWSS (Understanding Reading and Writing skills of school children). It uses the methodology of rule-based judgement. A specialist analysed the URAWSS test results. The findings were matched with the Decision Tree and Random Forest analysis conducted. The results demonstrated that using a Machine Learning (ML) technique detects better than using a traditional rulebased decision-making strategy. It is proposed that the model-based method of prediction using ML can also be devised using RTI (Response to Intervention).

Luz Rello et al in 2019., [6] worked for the prediction of dyslexia by using online screening tools. They conducted a web based gamified test for assessment of Spanish students. They designed the exercise by language skills and working memory. They used machine learning algorithm to train the model from human computer interaction. Random Forest is used because of the non-linearity of the data and achieved the accuracy of 89.2%. Screening test used in this was open source and also used 200,000 times in countries which speak Spanish.

An algorithm was developed by **Alex Frid and Zvia-Breznitz in 2016.,** [7] to identify people as Normal or readers which are Dyslexic by means of registered electroencephalogram (EEG) channels. For feature analysis and selection, the SVM is used. By popular vote, the ensemble of svm was used for classification and decision-making. Features such as maximum peak amplitude, Positive Field, and measurement of Spectral Flatness were extracted. The algorithm does not concentrate on dyslexia subtypes such as dyscalculia and dysgraphia, and instead deals with adult dyslexics. There was no record of the accuracy of this algorithm.

Perera et al in 2016., [8] worked through electroencephalogram-based research (EEG) to predict dyslexia. Electrodes have been used to track the brain's electrical activities. The support vector machine is used to extract features and classify dyslexics. pros and cons of the EEG method have been established and optimization methods have also been proposed to help predict dyslexia.

Rello et al in 2015., [9] developed a statistical model using eye-tracking systems to distinguish dyslexic readers from control readers. by Using trackers like eye trackers, movements of eyes are tracked. To construct the model, the Support Vector Machine binary classifier was used. Dyslexics were shown to exhibit longer fixations, smaller saccades, and increased regressions than non-dyslexic readers. The accuracy level of 80.18 % was achieved with 10-fold cross-validation. They suggested working on some other techniques for classification, such as learning from Neural Networks and Perceptron. Ballesteros, L. Rello Machine learning and eye tracking technologies are used to detect dyslexic readers. In: 12th Web for All Conference Proceedings. ACM

With the assistance of dyslexia metrics, **Al-Barhamtoshy et al in 2017.,** [10] proposed a model using classifier.The Gibson test of mental skills was used in the development of dyslexia measures. The supplied dataset is first pre-processed, analyzed, and interpreted using k-means classifiers, Artificial neural network and Fuzzy based classifier. These classifiers' outcomes help to categorize if a person is dyslexic or regular. By using three classifiers, an overall accuracy of 96

Zainuddin et al in 2016., [11] explains a classification technique for the identification of dyslexic children from the children which are normal applying EEG signals. The EEG signals were obtained and then adjusted for identification. This approach uses an enhanced version of the classifier k-nearest neighbor (KNN) to better distinguish between dyslexics and non-dyslexics.

Jain et al in 2009., [12] have applied Artificial Neural Network (ANN) perceptron model to classify reading difficulties by performing academic-based tests by remediation teachers.The model has an input layer with 11 units that correspond to different parts of the test and an output layer. Compared to other techniques, the technique requires less computing time and is easier to implement. It also provides robust and similar experimental findings on detection metrics including responsiveness, accuracy, and specificity.

2.2 Techniques for Image Processing

JainPłonski et al in 2017., [13] using multivariate classification techniques to analyze the disturbance of grey matter in brain images, examines the irregular anatomies of the dyslexic brain. MR brain scans of dyslexic youngsters from three different nations were used as input (French, German, and Polish). The photos are normalized in order to collect the properties that are needed for classification. According to the findings, dyslexia is defined by a regular curvature pattern in the left side of the brain with extra folds.. As compared to current systems, accuracy is higher.

A. El-Baz et al in 2008., [14] discussed ways of separating dyslexic and brain function. Quantitative study of the form of the gyrifications of Cerebral White Matter from three dimensional MRI is utilized for categorization. A person is categorized by measuring the amount of the extracted gyrifications as dyslexic or not. When comparing to the level-segment-based approach using segmentation, this quantitative classification outcome of this strategy is successful. The analysis can be expanded to explore other structures in the brain that can help to explain over a time span the shift in the dyslexic brain.In regards to CWM, gray matter can indeed be considered.

Tamboer et al in 2016., [15] researched morphometry based on voxel in dyslexia. Gray matter, volume alterations in white matter. Group differences in the white matter (WM) volume of local and total GM were analysed.

Dyslexia recognition and prediction is performed using databases of brain images. Photos were segmented using the Gaussian kernel and smoothed. Different cognitive features of dyslexia have been shown to have a high effect on anatomical changes dependent on dyslexia sub-types.

Feng et al in 2017., [16] during phonological processing in dyslexics, studied cerebellar functional integration and cerebellar activation. The brain images had been taken at Beijing Normal University using 3.0 T Siemens MRI (Magnetic Resonance Image) scanners. The input data gathered is preprocessed using SUIT (Software Process Management toolbox). Using a mathematical method, the images are then analysed. In children with dyslexia, unnaturally high bilateral cerebellum activation throughout orthographic activity has been observed. The findings indicate that during readings, the cerebellum was highly activated on both sides in dyslexics. They also suggested looking into the working brain connection between areas of the cerebral and some areas of cerebellar.

Mohamad et al in 2013., [17] investigated the relationship between fMRI brain imaging and 2D EEG topography in students which are dyslexic and non-dyslexic learners while they were studying. Filtering and feature extraction are performed on the obtained EEG signal. It examined whether EEG topography can be substituted to make prediction less costly and non-invasive in terms of MRI and PET. The left side of the brain of the brain is activated in normal children, the right hemisphere is activated in dyslexic competent children, and both the left and right hemispheres are activated in dyslexic weak children. As a result, differences in hemispheric stimulation between dyslexic and learners that are non-dyslexics were discovered.

In dyslexic learners, **Morken et al in 2017.,** [18] examined the cortical integration of brain photos. Three separate literacy classes were among the participants.The dynamic causal model (DCM) is designed to quantify the

connection measurements. These parameters are subsequently examined using the ANOVA analytic approach. The findings revealed that during a transition into pre-emergent and emergent literacy levels, the dyslexics have shown a delay.

Author	Research Objective	Model or Algorithm	Pros	Cons	Future Work
Zvia, Breznitz, Alex Frid (2012)	Using ERP signals, differentiate dyslexic readers from healthy readers using the support vector machine method.	SVM	Easy model to identify dyslexia	The subcategories of dyslexia, such as dyscalculia and dysgraphia, are not addressed.	Features associated with dyslexia types can be identified and utilized to diagnose dyslexia subtypes.
Perera et al.(2016)	Dyslexia categorization based on EEG analysis	SVM	Identified the benefits and drawbacks of utilizing EEG to predict dyslexia.	Optimization approaches that might help with prediction isn't taken into account.	Mirror prox method and other optimization approaches can be utilized.
Iwabuchi(2017)	Assessment of reading and writing problems using machine learning	Random Forest and Decision Tree	When compared to a rule-based method, prediction accuracy is higher.	A reaction to Intervention model may be used to improve prediction.	For prediction, a rule-based decision method combined with machine learning can be utilized.
Rello et al.(2019)	Measures of eye movement to predict dyslexia	SVM	When compared to some other systems, the level of accuracy (80.18 percent) is high.	Level of accuracy can be enhanced	It is possible to improve feature extraction.
Palacios (2019)	Dyslexia diagnosis based on low-quality data	Induction method for fuzzy unordered rules	Assessment by parents and teachers benefits in improving accuracy.	The amount of time required to categorize a dataset using FURIA is huge.	The use of a distributed approach helps speed up the execution process.
A. El-Baz (2008)	Using MRI, a CAD system for early identification of dyslexic brains has been developed.	QSA(Quantitative shape analysis)	This method is finer than the level segment method.	The volume of gray matter is not taken into account.	Extend the research to learn more about how the dyslexic brain changes over time.
Tamboer (2016)	Machine learning is used to classify neuroimaging images to determine if they are dyslexic or not.	For classification, use a Gaussian Kernel.	Dyslexia subtypes are investigated.	The severity and kinds of dyslexia are not taken into account while determining the classification.	For a better identification of Subtypes, more characteristics can be retrieved.
Cui et al.(2016)	Machine learning techniques were used to investigate the relationship between white matter interconnection and developmental dyslexia.	SVM	There is an accuracy of 83.61 percent.	Only the white matter areas of the brain were examined.	The gray matter and white matter of the brain both be studied.

Table 2.1: Literature Review Summary

Author	Research Objective	Model or Algorithm	Pros	Cons	Future Work
Morken (2017)	A study of dyslexic reading across literacy categories and growth.	Analysis using ANOVA	Dyslexics' movements have been found to be delayed.	A smaller number of individuals were examined.	On huge datasets, check for classifier accuracy.
Mohamad (2013)	Neurological approaches for diagnosing dyslexia in children are being researched.	Uses statistical model	There was a difference in hemispheric activation.	Only a few characteristics were observed.	The collection of different classifiers
Al-Barhamtoshy (2017)	Dyslexia diagnosis based on computational analysis	ANN, K-means and fuzzy classifier	By using three classifiers, an average accuracy level of 96 percent was attained.	It would be more beneficial to use EEG signals when reading or writing.	To detect dyslexics, record EEG patterns while reading.
Zainuddin (2016)	EEG for distinguishing competent dyslexic children from non-dyslexic youngsters	KNN	When compared to previous systems, classification error is lower.	Before using KNN, you may standardize your data.	For better accuracy, look at different clustering models.
Jain et al.(2009)	Learning disability diagnosis through computation	Artificial neural network based on perceptrons	Simplicity and sophistication are reduced.	Machine learning and evaluation by a specific teacher can be combined.	The collection of machine learning algorithms
Płonski (2017)	Identification of developmental dyslexia by neuroanatomical studies	A multiparameter machine learning technique	The left hemisphere of the brain has a regular curvature structure with additional folds which was detected as a feature of dyslexia	Other areas of the brain aren't taken into account.	The amount of white matter and gray matter in the brain can be utilized to identify dyslexics.

Table 2.2: Literature Review Summary

2.3 Summary

This chapter presented an overview of the techniques presented for the identification of dyslexia through different methods. Some researchers have used data mining techniques, some used machine learning algorithms and some detected dyslexia by MRI scans, CT scans or eye ball tracking system. many different type of datasets are used by different researchers. in some areas language is also a major factor in development of dyslexia. very few people have researched on the different types of dyslexia because it is very challenging. none of them identified the different levels of within dyslexics.

CHAPTER 3

METHODOLOGY

In the preceding chapter, significant contributions in the field of dyslexia identification both in medical and computer science in the last two decades were discussed. In the previous studies, researcher considered few or too many factors while detecting dyslexia, however that is not enough for effective identifying this type of difficulty. As few features cannot detect dyslexia, the detection includes many important factors because it is very difficult to detect dyslexia by just 2 or 3 factors.

Factors that are considered while detection of dyslexia are spelling score, reading score, memory score, verbal score, writing speed, accuracy score, punctuation score, visual score, ability score, vocabulary score, difficulty with handwriting, confidence level, avoidance, lack of motivation and comprehension score etc. These are the factors considered but in this research we have taken into account only important features.

Dyslexia is a disease that is very common in our society but people often could not find the difficulty through which the child is suffering. Delay in treatment can lead to failure of a student. Students leave schools, indulge in crimes, and join the company of bad people. This all happens because of negligence of our parents and teachers. This study helps to find out the main causes behind slow learner and most of the important factors that are present in any dyslexic students. So where parents and teachers have to put more

effort on student having symptoms of this difficulty. Although dyslexia is not a type of disease which is fully treated but parents and professionals can give confidence and ability to rise and shine in the field of interests of students.

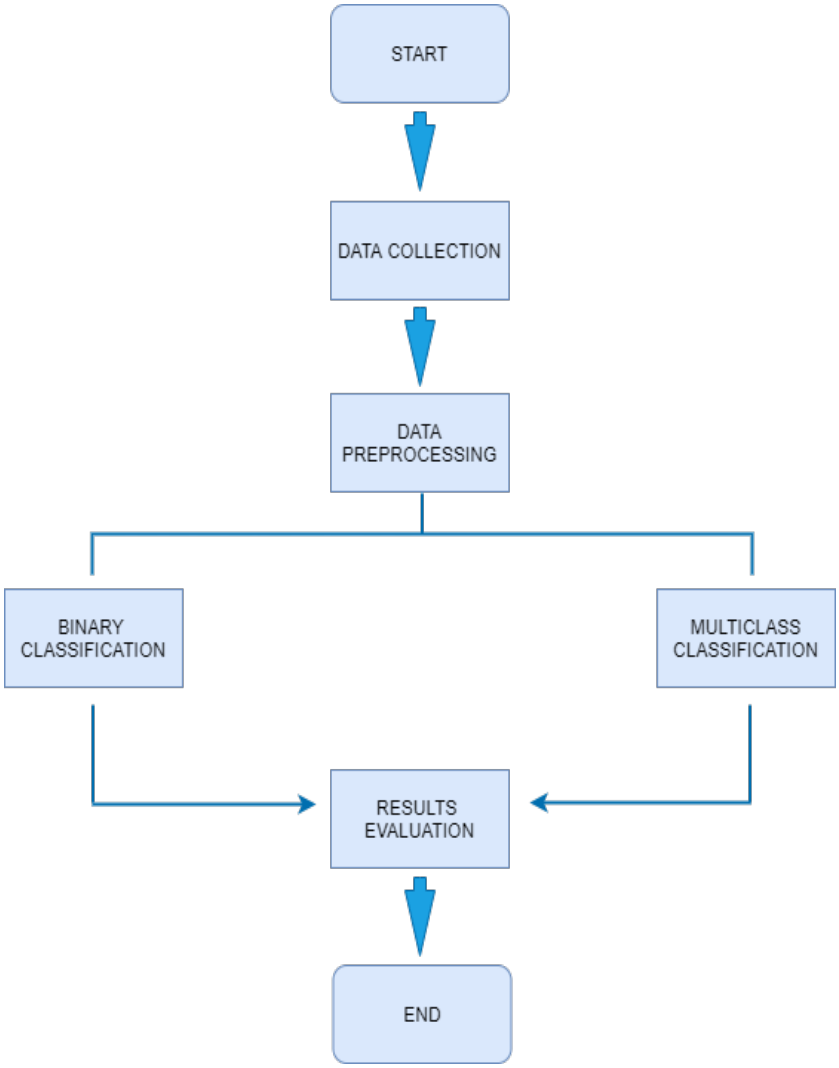


Figure 3.1: An Overview of Key Processing Steps in Classification.

We now present the details of the proposed methodology for the classification of dyslexia from the challenging dataset. We first introduced the dataset employed in our study followed by the details of pre-processing, data collection and detection of dyslexia through different algorithms. Binary classification and multiclass classification is performed separately and than results are evaluated.

An overview of the key steps of classification is presented in fig: 3.1 and steps performed while clustering are shown in fig: 3.2 while each of these steps

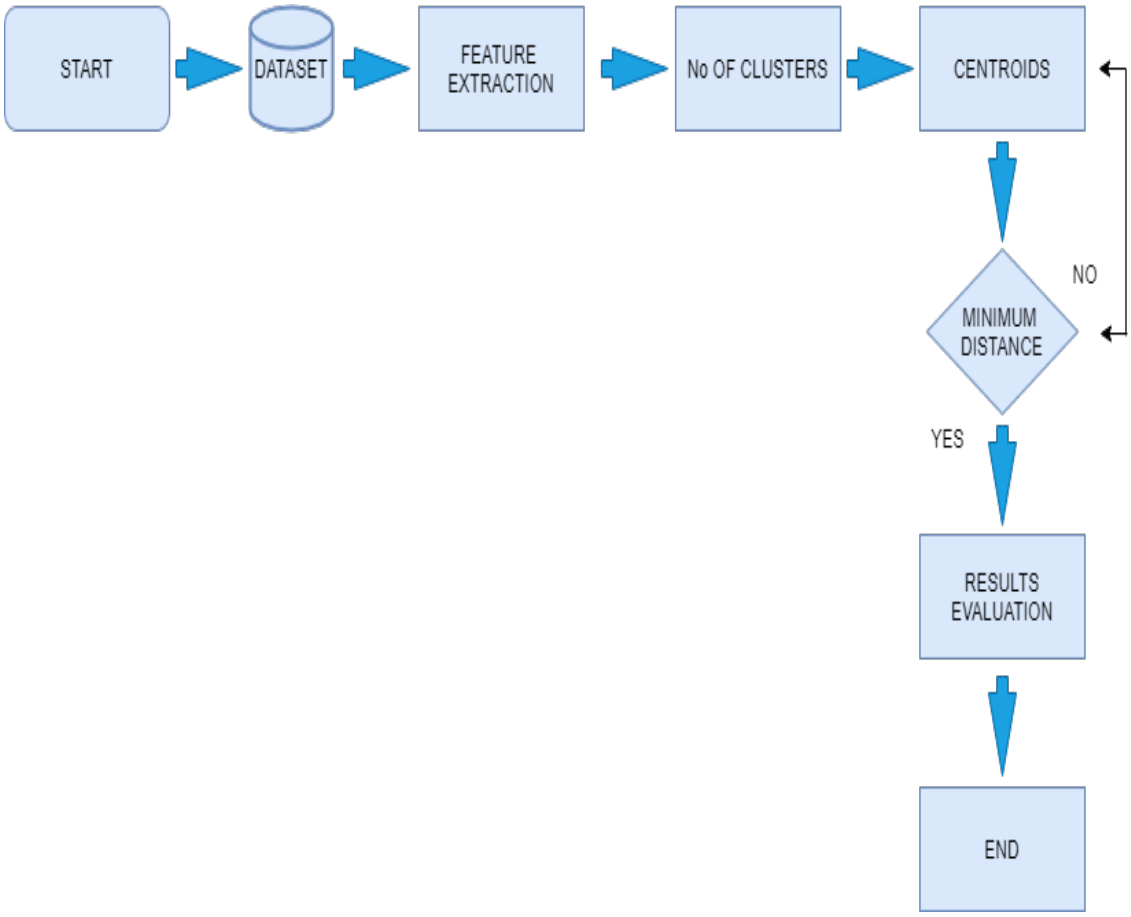


Figure 3.2: An Overview of Key Processing Steps in Clustering.

is discussed in detail in the subsequent sections.

3.1 Dataset

Dataset used in this research is provided by Bazeecha trust of IDEAS foundation Islamabad. The dataset used in this research is of dyslexic patients. It consists of 8 attributes, 2 labels and 200 records. Some attributes were categorical but in preprocessing they were converted in numerical. There were also some null values but also they were replaced with non-numerical values in preprocessing step. This dataset contains records of both girls and boys, as gender cannot play any role in detection and analysis. Age of these students range from 7 to 10 years. The dataset used in this research consist of students

roll no, attributes, and labels. Following attributes are used in this research:

- Spelling score
- Reading score
- Visual score
- Vocabulary score
- Memory score
- Writing score
- Punctuation score
- Comprehension score

Spelling and reading scores range from 0 and 1. Visual, vocabulary and writing scores were categorized as average, below average or above average but after preprocessing these replaced. Comprehension, memory and punctuation scores were initially categorized as starter or non-starter. There are two column of labels, first for binary classification labeled as 0 and 1. Second for multiclass classification labeled as non-dyslexic, mild, moderate and severe. This column represents severity of dyslexic child.

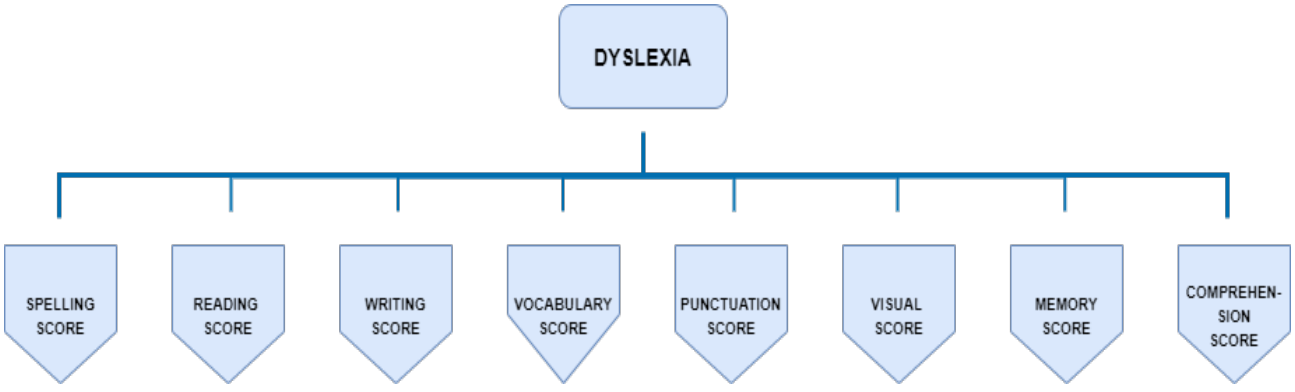


Figure 3.3: Modules for Identification

3.2 Data Collection

Bazeecha trust is an organization for special children, they assist youngsters in overcoming early trauma and regaining self-acceptance, dignity, and faith in humanity. Their goal is to offer housing, excellent education, and long-term livelihood possibilities to the poor by promoting self-reliance and financial independence in them so that they may contribute to their communities as productive members.

Bazeecha Trust of IDEAS foundation Islamabad provided the dataset used in this research. They are using different testing batteries for basic assessments of students having symptoms of dyslexia. In this trust there are professionals who assists and educate dyslexic students. They first take preliminary tests required for initial assessment for either the student is dyslexic or not. These tests contain different questions based on skills like reading, writing and sounds. Students need to answer all the questions in the questionnaire and on the basis of those, the groups will be made based on the attributes.

3.3 Pre-processing

The primary objective of pre-processing is to normalize the dataset. Normalization means re-scaling of real numeric attributes to range between 0 to 1. Data normalization is used in machine learning to make our model more accurate because by this the model will converge to better weights. First, normalization is done on the dataset for re-scaling the numeric values to the range from 0 to 1.

Mostly machine learning algorithms only take numerical data. They only take numerical data as input. Some categorical values are replaced. Categorical values include categories or different type of levels that are non-numerical and cannot be processed as it is. So for this purpose this type of data should

be preprocessed for further processing.

3.4 Methodology

In this research, we have used both supervised and un-supervised learning algorithms. Supervised learning algorithms are those in which labels are given for prediction but in un-supervised learning labels or ground truths are not mentioned in dataset. In supervised learning, we have implemented binary and multiclass classification and for unsupervised learning, we have implemented clustering. We have classified the students into dyslexic and non-dyslexic classes using binary classification. Multiclass classification is used for classifying different classes for identification of different severity levels amongst dyslexics. There are different level of severities but here there are 3 types which are mild, moderate and severe. After classification, we have done subgroup analysis by applying clustering on dyslexic students to group them on the basis of similarities.

We have classified the data based on dyslexic and non-dyslexic classes using ANN and multiclass classification using SVM , and for suspected dyslexic classes we have further assessed the groups using kmeans clustering. The data is splitted in the ratio of 80:20, 80% data is used for training and 20% data is used for testing. We have done clustering for the purpose of subgroup analysis to find similarities. In classification we have explored different algorithms and found that for binary classification neural network is giving better results and is more suitable for binary classification as a machine learning approach. For multiclass classification, svm is used because for identifying more than 2 classes it works perfectly.

Multiclass classification is done for detecting different classes for individual rehabilitation. This is essential for giving adequate treatment to each learner. Clustering is used for grouping based on similar attributes (mentioned in the dataset). clustering is done for identification of hidden patterns and sim-

ilarities. Those similarities could be severity level, different types of dyslexia or may be some other type of difficulty. This is done for the purpose of analysis of groups or clusters made after applying kmeans.

In this model, identification of similar dyslexic children is done on the basis of some attributes because there are different types of dyslexia and also level of severity is different. That is why treatment, therapies and rehabilitation of different children must not be same. Previously the work that has been done in this regard is purely limited to the diagnosis of dyslexia however, there is still a need of research in identifying the types, severity levels as well as analysis on hidden similarities based on similar types of dyslexia to help in understanding and channeling the process of treatment and rehabilitation according to the severity or other related factors.

3.5 Summary

This chapter introduced the proposed dyslexia classification and analysis technique along with the pre-processing techniques investigated. Different techniques for preparing the data for processing is applied. By using these, the model has given better and accurate results. Classification is done using neural network and clustering is done using k-means algorithm. Training and testing is done on the same dataset as it was splited by the algorithm. The model is trained on the splitted data which is dedicated for training and then tested on the rest of dataset. The next chapter presents the details of the experimental study carried out to validate the proposed methods.

CHAPTER 4

ANALYSIS & RESULTS

This chapter presents the details of the different experiments with analysis and results. We first introduce the experimental protocol followed by the identification performance of different models. In this research, binary classification, multiclass classification and clustering is used. Multiclass classification is used for identifying different levels or severity and clustering algorithm is used for the purpose of subgroup analysis. Subgroup analysis is important for detection of unseen patterns and deep insights.

4.1 Experimental Protocol

The dataset for detection and analysis of dyslexia is collected from Bazeecha trust Islamabad. The dataset contains 200 records of students with 8 attributes and two labels. This outcome will tell that the student is dyslexic or not and second label will tell the level of severities. Some of the attributes are spelling score, reading score and vocabulary score etc. In this dataset 0 represents non-dyslexics and 1 represents dyslexics. Threshold for dyslexics was set to 62%. Students having total percentage greater than this is considered as non-dyslexics and students having percentage less than 62% is considered as dyslexics.

Initially the dataset is divided into training data and test data. The distribution of data is done in the ratio of 80:20. 80% data is used for training

purpose and remaining 20% is used in testing purpose. First for binary classification, neural network is used having 3 hidden layers. The activation functions that are used in this research are sigmoid and relu. Than for multiclass classification, svm is applied using polynomial kernel. In multiclass classification, label encoder is used for encoding the labels initially named as non-dyslexic, mild, moderate and severe. Principle Component Analysis (PCA) is used in clustering for dimensionality reduction to visualize clusters in an appropriate manner.

There are two types of machine learning algorithms: supervise and unsupervised learning. Learning from pre-trained data having labels is known as supervised learning. Learning through the data without any training and labels is known as un-supervised learning. The most commonly used machine learning techniques are neural networks, support vector machine, decision tree, k-means and logistic regression.

4.2 Classification

1. Binary Classification

2. Multiclass Classification

4.2.1 Binary Classification

It is a classification method or task in which a set of data is divided into two groups. It's mainly a prediction as to which of two categories the object belongs to. Many Algorithms are used for this type of classification like ANN, Decision tree and svm. For classification purpose we have analyzed different algorithms like ANN, SVM and decision tree. However ANN gives best results as compared to these two classifiers. Classification report of binary classification is shown in table: 4.1

Artificial Neural networks imitate the brain's ability to comprehend infor-

mation and make various predictions based on that knowledge. It is basically a group of neurons grouped together. Architecture of ANN is shown in fig: 4.1. Artificial Neural Networks are made up of three layers:

1. Input
2. Hidden
3. output

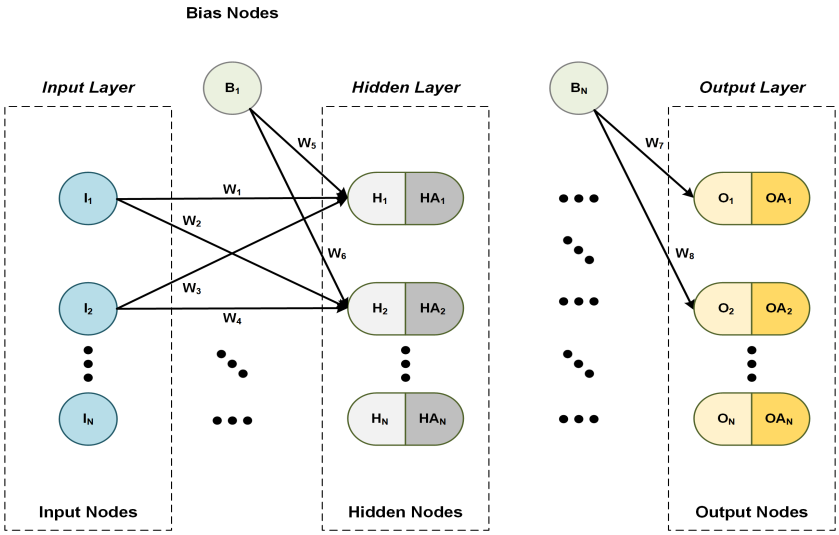


Figure 4.1: General architecture of a neural network [19]

The unprocessed features are in the input layer, the highly connected neurons are in the hidden layer, and the prediction is in the output layer. For accurate prediction, neurons are fine-tuned. Every input is first multiplied by a weight. Then, with a bias b, all of the weighted inputs are summed together. Finally, an activation function is applied to the sum.

The activation function is used to convert an infinite input to a desirable, predictable output. The sigmoid function is a popular activation function. Sigmoid function gives output only in the range between 0 and 1. Large negative numbers convert into 0 and large positive numbers converts into 1.

Right now, the ReLU is the most widely used activation function in machine learning. Since then, it's been employed in nearly all convolutional neural networks and deep learning systems. Its range is from 0 to infinity and for negative values it becomes 0. Model cannot properly train from this behavior of relu.

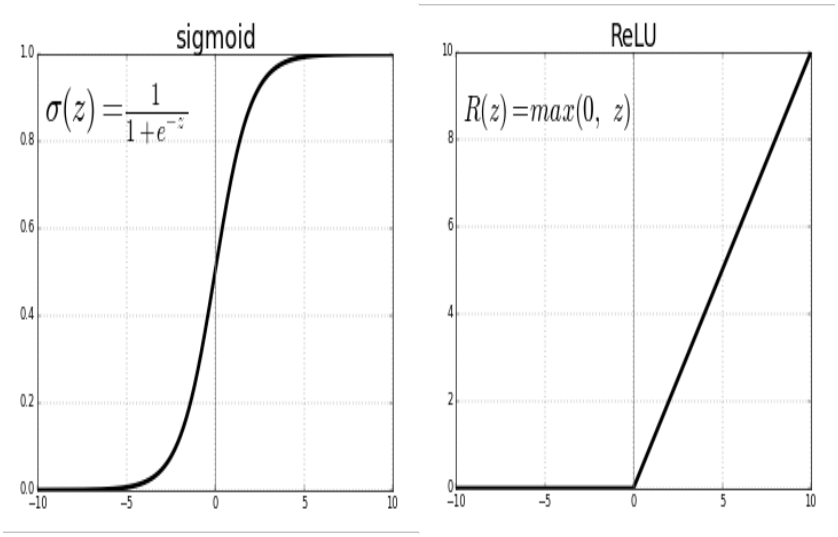


Figure 4.2: Sigmoid and relu

[20]

4.2.2 Multiclass Classification

Multiclass classification done for the purpose to make different classes for severity levels to give individual rehabilitation. Severity levels on which multiclass classification is done are mild, moderate and severe. As this is very important because there are multiple of factors involved in detection. Adequate treatment and therapies are most important for each individual.

The task of classifying elements into various classes is known as multi-class classification. It is not limited to any number of classes, unlike binary. There are several classes for the response variable to be categorized in, and hence it is referred to as a Multi-class classification. For this classification we have used SVM because it performed well and gave good results.

SVM is a supervised machine learning method that may be used to help with classification and regression issues. Its goal is to discover the optimal boundary between the possible outputs. SVM performs complicated data transformations according on the kernel function chosen, and then seeks to optimize the separation boundaries between your data points based on the labels or classes you've established. Classification report by using svm is shown in table:4.2.

4.3 Clustering

For the purpose of subgroup analysis k-means algorithm is used as it is popular and performs well with unlabelled data. K-means clustering is an unsupervised learning technique for identifying unique data groups having similar characteristics. It works with data for which the class labels are unknown. The number K in the name indicates to the number of groups that were generated. It is used in a variety of corporate applications. 7 clusters were made because of different level of similarities. As this is very important for treating different levels of students differently.

Subgroup analysis is very important because there are many underlying hidden similarities and it is very necessary to identify these. These can be deeply analyzed by professionals in future. These similarities can be different severity level, types of dyslexia or any other difficulty that cannot be identified without professionals. It is very important to give rehabilitation and treatment to each individual according to the level of difficulty otherwise all the students would have given the same treatment.

Different groups are identified on the basis of similarities using clustering algorithm. The clusters made are shown in fig: 4.3 up to fig: 4.7. These clusters can further be used for detailed analysis.

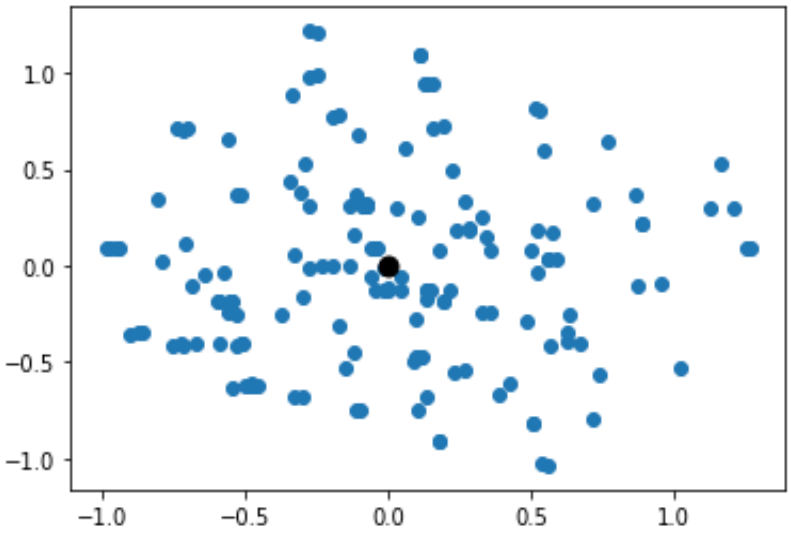


Figure 4.3: K-means plot with 1 Cluster

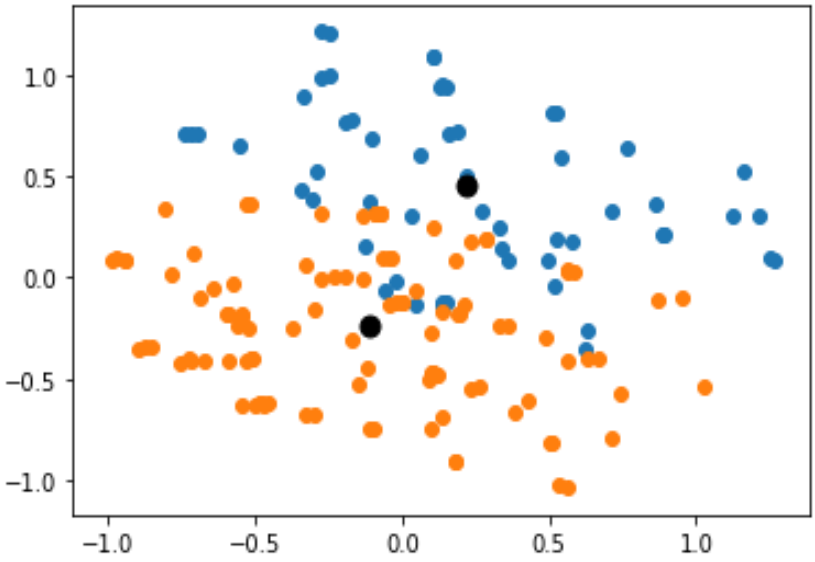


Figure 4.4: K-means plot with 2 Clusters

These graphs shows the no of clusters made by using kmeans algorithm. Also different colors are used to represent clusters and black dots shows the centroid of each cluster. By these clusters it is identified that different sub-groups exists within dyslexia. These can be different levels of severity other

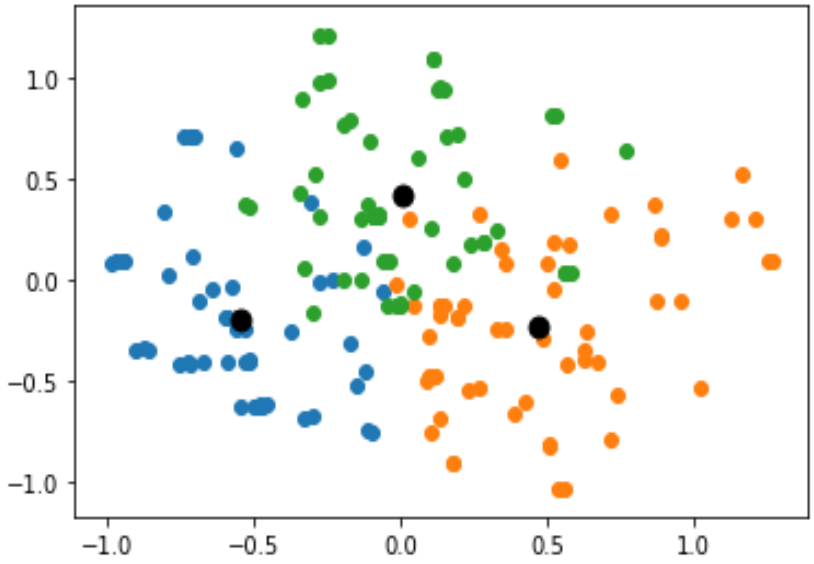


Figure 4.5: K-means plot with 3 Clusters

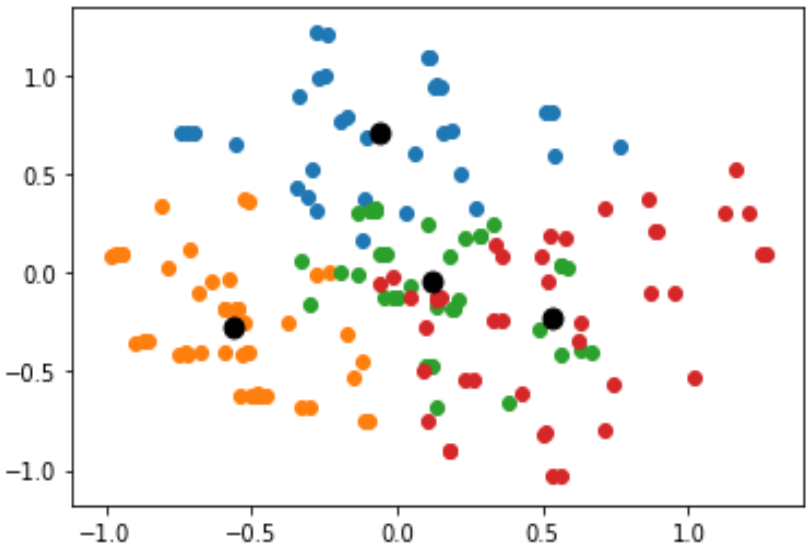


Figure 4.6: K-means plot with 4 Clusters

than mild, moderate and severe, different types of dyslexia, any other difficulties or groups made due to some underlying similarities. These groups are very important because on the basis of these, adequate individual treatment and therapies must be given.

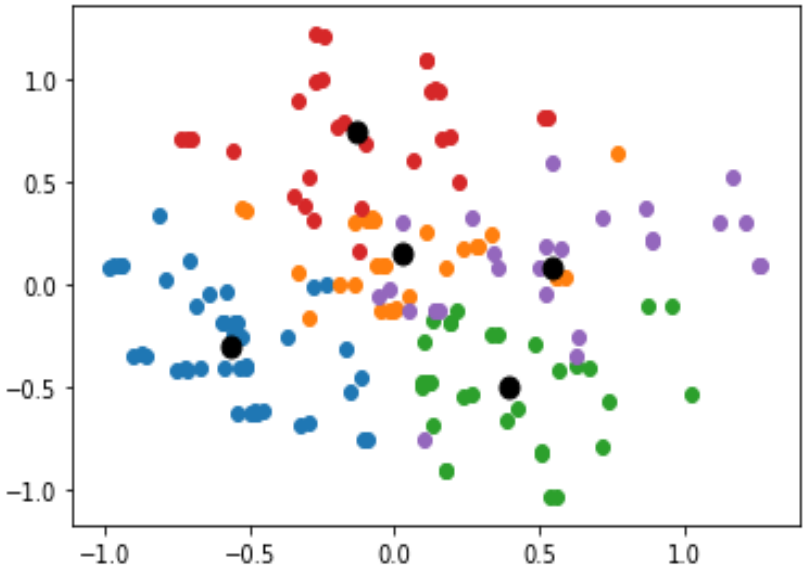


Figure 4.7: K-means plot with 5 Clusters

The main purpose of clustering is to identify different subgroups exists within dyslexia so, therapists or doctors can treat every student according to their specific type of treatment. Detail analysis on these subgroups can be further analyzed by other researchers.

4.4 Evaluation Metrics

The results of experiments for detecting this disease are discussed in this section. Machine learning algorithm for classification purpose is applied. For classification purpose neural network is applied on the real data. It is used because it is popular and usually gives better results.

4.4.1 Precision, recall and f1-score

The model is evaluated by evaluation metrics like precision, recall and f1-score. Data used in this research is imbalanced so for this purpose accuracy is not considered as a good evaluation metrics. The results of experiment show that, for binary classification neural network and for multiclass classification

support vector machine has performed very well on this dataset with these features. The classification report for binary classification is as shown in table: 4.1 and the classification report for multiclass classification is shown in table: 4.2

	precision	recall	f1-score
0	0.5	0.38	0.43
1	0.85	0.91	0.88
accuracy			0.8
macro avg	0.68	0.64	0.65
weighted avg	0.78	0.8	0.79
weighted avg	0.9	0.88	0.88

Table 4.1: Classification Report of Binary classification (ANN)

	precision	recall	f1-score
0	0.57	1	0.73
1	0.89	0.89	0.89
2	0.96	0.85	0.9
accuracy			0.88
macro avg	0.81	0.91	0.84
weighted avg	0.9	0.88	0.88

Table 4.2: Classification Report of Multiclass Classification (SVM)

4.5 Evaluation metric of clustering

4.5.1 Elbow Method

Based on the sum of squared distance (SSE) between data points and their assigned clusters' centroid, the elbow approach provides us an estimate of what a suitable k number of clusters might be. At the point when SSE begins to flatten out and form an elbow, we choose k. By using the dyslexia dataset, we have calculated SSE for various k values to observe the curve shown in fig: 4.8. The graph below shows that k=3 is a good choice and shows an elbow.

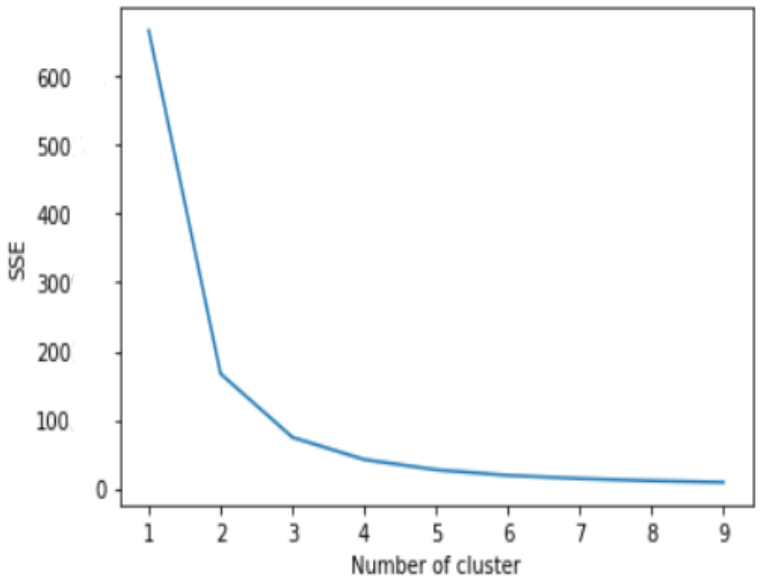


Figure 4.8: Elbow Method(SSE)

4.5.2 Silhouette coefficient

The silhouette coefficient, often known as the silhouette score, is a measure for determining how successful a clustering approach is. Its value is between -1 and 1. A greater Silhouette Coefficient means the object is well matched to its own cluster but not to nearby clusters. A model with better-defined clusters has a higher Silhouette Coefficient score.

The Silhouette Coefficient for n clusters=2 is the greatest. As a result, the ideal number of clusters should be two. Despite the fact that n clusters=2 has the greatest Silhouette Coefficient, we believe n clusters=3 is the optimal number of clusters.Choosing optimal no. of the cluster will depend on the type of dataset and the problem we are trying to solve.

4.6 Related Work

Many researchers have worked related to this research. Some of them are shown in Table 2.2. Here is comparison between our research with the ones with the detection of dyslexia using machine learning algorithms. As mentioned above that data used in this research is imbalance so we evaluated our results using precision, recall and f1-score. So we compared these metrics with others and the comparison is shown in table: 4.3 and also in the form of graph in fig: 4.9. In some studies the data in datasets only contains records of 40 to 50 students, so because of less amount of samples model cannot train properly and gives results like accuracy either too low or too high.

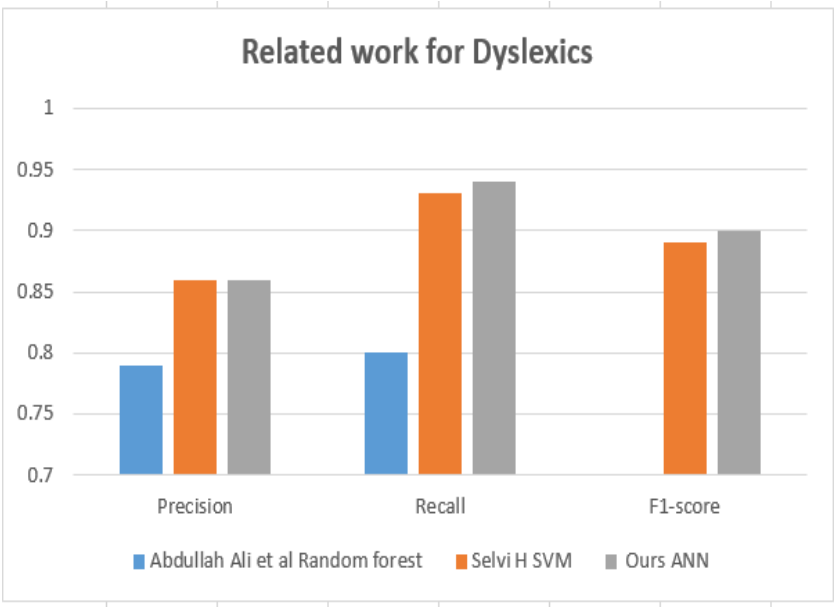


Figure 4.9: Related Work

Related work for Dyslexics				
Author	**Algorithm**	**Precision**	**Recall**	**F1-score**
Abdullah Ali et al	Random forest	0.79	0.80	0.79
Selvi H	SVM	0.86	0.93	0.89
Ours	ANN	0.86	0.94	0.9

Table 4.3: Comparison of Precision, Recall and F1-score by Related work

CHAPTER 5

CONCLUSION & FUTURE WORK

5.1 Significance of the study

Dyslexia is a disease which is very common in our society but no one tries to find the root cause. Delay in treatment can lead to failure of a student. Students leave schools and indulge in crimes and join the company of bad people. This all happens because of negligence of our parents and teachers. This study helps to find out the main causes of slow learner and which factors are there in any dyslexic students. So where parents and teachers have to put more effort on student having symptoms of this difficulty.

5.2 Scope and limitation

The scope of the proposed model is to identify the groups within dyslexic children based on different attributes to provide similar rehabilitation and treatment to child with similar dyslexic case, as mentioned previously that the dyslexia cannot be determine based on any single attribute as well as all the associated problems and severity of similar types of dyslexia cannot also be the same. So, considering this into account the treatment, rehabilitation and intervention of similar type of dyslexia cannot be the same. So, the purpose of this research is to identify the groups within dyslexic children.

In addition to that for the purpose of detection of 2 or more classes, classification techniques are used and for the purpose of grouping clustering

algorithms are used. The research used clustering based on following attributes spelling score, reading score, visual score, vocabulary score, comprehension score, memory score, writing score and punctuation score.

5.3 Future Work

In this research we have strictly focused on identifying different classes and grouping based on similar attributes in dyslexic cases. In addition to that the research also not focused on diagnosis and its techniques in terms of dyslexia, also the research do not identified the type of severity. Future researchers can work on detailed subgroup analysis for identifying levels of severity for better treatment and individual therapies.

In our further study on this subject, we intend to extend the analysis to other relevant problems like the classification of writing styles and dating. Furthermore, the current study revealed that pre-processing is a critical step in analyzing such documents and further investigating different pre-processing techniques could indeed be an interesting study. In addition to standard pre-trained models, relatively shallower networks can also be trained from scratch to study the performance evolution. It is expected that the findings of this study would be helpful for the pattern classification community in general and the handwriting analysis community in particular.

REFERENCES

[1] D. Times, 12 million children in pakistan have dyslexia.
URL http://dailytimes.com.pk/120033/12-million-children-in-pakistan-have-dysl

[2] I. Cimermanová, Teaching english as a foreign language to dyslexic learners, Teaching Foreign Languages to Learners with Special Educational Needs (2015) 39–62.

[3] B. A. Shaywitz, S. E. Shaywitz, B. A. Blachman, K. R. Pugh, R. K. Fulbright, P. Skudlarski, W. E. Mencl, R. T. Constable, J. M. Holahan, K. E. Marchione, et al., Development of left occipitotemporal systems for skilled reading in children after a phonologically-based intervention, Biological psychiatry 55 (9) (2004) 926–933.

[4] I. S. Isa, W. N. S. Rahimi, S. A. Ramlan, S. N. Sulaiman, Automated detection of dyslexia symptom based on handwriting image for primary school children, Procedia Computer Science 163 (2019) 440–449.

[5] M. Iwabuchi, R. Hirabayashi, K. Nakamura, N. K. Dim, Machine learning based evaluation of reading and writing difficulties., Studies in Health Technology and Informatics 242 (2017) 1001–1004.

[6] L. Rello, R. Baeza-Yates, A. Ali, J. P. Bigham, M. Serra, Predicting risk of dyslexia with an online gamified test, arXiv preprint arXiv:1906.03168 (2019).

[7] A. Frid, Z. Breznitz, An svm based algorithm for analysis and discrimination of dyslexic readers from regular readers using erps, in: 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, IEEE, 2012, pp. 1–4.

[8] H. Perera, M. F. Shiratuddin, K. W. Wong, A review of electroencephalogram-based analysis and classification frameworks for dyslexia, in: International Conference on Neural Information Processing, Springer, 2016, pp. 626–635.

[9] L. Rello, M. Ballesteros, Detecting readers with dyslexia using machine learning with eye tracking measures, in: Proceedings of the 12th Web for All Conference, 2015, pp. 1–8.

[10] H. M. Al-Barhamtoshy, D. M. Motaweh, Diagnosis of dyslexia using computation analysis, in: 2017 International Conference on Informatics, Health & Technology (ICIHT), IEEE, 2017, pp. 1–7.

[11] A. Zainuddin, K. Y. Lee, W. Mansor, Z. Mahmoodin, Optimized knn classify rule for eeg based differentiation between capable dyslexic and normal children, in: 2016 IEEE EMBS conference on biomedical engineering and sciences (IECBES), IEEE, 2016, pp. 685–688.

[12] K. Jain, P. Manghirmalani, J. Dongardive, S. Abraham, Computational diagnosis of learning disability, International Journal of Recent Trends in Engineering 2 (3) (2009) 64.

[13] P. Płoński, W. Gradkowski, I. Altarelli, K. Monzalvo, M. van Ermingen-Marbach, M. Grande, S. Heim, A. Marchewka, P. Bogorodzki, F. Ramus, et al., Multi-parameter machine learning approach to the neuroanatomical basis of developmental dyslexia, Human Brain Mapping 38 (2) (2017) 900–908.

[14] A. El-Baz, M. Casanova, G. Gimel'farb, M. Mott, A. Switala, E. Vanbogaert, R. McCracken, A new cad system for early diagnosis of dyslexic brains, in: 2008 15th IEEE International Conference on Image Processing, IEEE, 2008, pp. 1820–1823.

[15] P. Tamboer, H. Vorst, S. Ghebreab, H. Scholte, Machine learning and dyslexia: Classification of individual structural neuro-imaging scans of students with and without dyslexia, NeuroImage: Clinical 11 (2016) 508–514.

[16] X. Feng, L. Li, M. Zhang, X. Yang, M. Tian, W. Xie, Y. Lu, L. Liu, N. N. Bélanger, X. Meng, et al., Dyslexic children show atypical cerebellar activation and cerebro-cerebellar functional connectivity in orthographic and phonological processing, The Cerebellum 16 (2) (2017) 496–507.

[17] S. Mohamad, W. Mansor, K. Y. Lee, Review of neurological techniques of diagnosing dyslexia in children, in: 2013 IEEE 3rd International Conference on System Engineering and Technology, IEEE, 2013, pp. 389–393.

[18] F. Morken, T. Helland, K. Hugdahl, K. Specht, Reading in dyslexia across literacy development: A longitudinal study of effective connectivity, Neuroimage 144 (2017) 92–100.

[19] C. Szegedy, A. Toshev, D. Erhan, Deep neural networks for object detection (2013).

[20] A. Bonner, Activation functions in neural networks (2017).
URL https://towardsdatascience.com/activation-functions-neural-networks-1cbd9

Turnitin Originality Report

Processed on: 16-Sep-2021 10:17 PKT
ID: 1649678800
Word Count: 7723
Submitted: 1

thesis By Ammara Minhas

Similarity Index	Similarity by Source	
17%	Internet Sources:	9%
	Publications:	8%
	Student Papers:	9%

3% match (publications)
A. Jothi Prabha, R. Bhargavi. "Chapter 3 Prediction of Dyslexia Using Machine Learning—A Research Travelogue", Springer Science and Business Media LLC, 2019

2% match (student papers from 01-Feb-2021)
Submitted to University of Limerick on 2021-02-01

2% match (student papers from 30-Nov-2020)
Submitted to University of Limerick on 2020-11-30

2% match (publications)
"Pattern Recognition and Artificial Intelligence", Springer Science and Business Media LLC, 2021

1% match (student papers from 25-Oct-2020)
Submitted to University of Limerick on 2020-10-25

1% match (student papers from 09-Dec-2019)
Submitted to Asia Pacific University College of Technology and Innovation (UCTI) on 2019-12-09

1% match (student papers from 14-Dec-2017)
Submitted to Asia Pacific University College of Technology and Innovation (UCTI) on 2017-12-14

1% match (student papers from 10-Jun-2021)
Submitted to Victorian Institute of Technology on 2021-06-10

< 1% match (student papers from 14-Sep-2021)