A Mini Project Report on

Machine Learning Approach for Early Identification of Dyslexia in Children

Submitted to the Department of Computer Science & Engineering, GNITS in the partial fulfillment of the academic requirement for the award of B. Tech (CSE) under JNTUH, Hyderabad

By

P.Sathwika	**(21251A0521)**
G.Manaswini	**(21251A0542)**
P.Varsha	**(21251A0565)**
K.Samhita	**(22255A0501)**

Under the guidance of

Mrs. K. Sindhura
Assistant Professor

Department of Computer Science & Engineering
G. Narayanamma Institute of Technology & Science
(Autonomous) (For Women)

Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad
Accredited by NBA & NAAC, an ISO 9001:2015 certified Institution
Shaikpet, Hyderabad-500104
July, 2024

A Mini Project Report on

Machine Learning Approach for Early Identification of Dyslexia in Children

Submitted to the Department of Computer Science & Engineering, GNITS in the partial fulfillment of the academic requirement for the award of B. Tech (CSE) under JNTUH, Hyderabad

By

P.Sathwika	**(21251A0521)**
G.Manaswini	**(21251A0542)**
P.Varsha	**(21251A0565)**
K.Samhita	**(22255A0501)**

Under the guidance of

Mrs. K. Sindhura
Assistant Professor

Department of Computer Science & Engineering
G. Narayanamma Institute of Technology & Science
(Autonomous) (For Women)

Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad
Accredited by NBA & NAAC, an ISO 9001:2015 certified Institution
Shaikpet, Hyderabad-500104
July, 2024

Department of Computer Science & Engineering
G. Narayanamma Institute of Technology & Science

(Autonomous) (For Women)
Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad
Accredited by NBA & NAAC, an ISO 9001:2015 certified Institution
Shaikpet, Hyderabad-500104

Certificate

This is to certify that the Mini Project report on "**Machine Learning Approach for Early Identification of Dyslexia in Children**" is a bonafide work carried out **P.Sathwika (21251A0521), G.Manaswini (21251A0542), P.Varsha (21251A0565), K.Samhita (22255A0501)** in the partial fulfillment for the award of B. Tech degree in Computer Science & Engineering, G. Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, affiliated to Jawaharlal Nehru Technological University, Hyderabad under our guidance and supervision.

The results embodied in the Mini project work have not been submitted to any other University or Institute for the award of any degree or diploma.

Internal Guide
Mrs. K. Sindhura
Assistant Professor
Department of CSE

Head of the Department
Dr. A. Sharada
Professor and Head
Department of CSE

External Examiner

To whom so ever it may concern

This is to certify that **P.Sathwika (21251A0521), G.Manaswini (21251A0542), P.Varsha (21251A0565), K.Samhita (22255A0501)** IV B.Tech of CSE Department, G. Narayanamma Institute of Technology & Science, Hyderabad, has successfully completed project work in **"Research Center of Center for Assistive Technology "** CSE.

The project titled **"Machine Learning Approach for Early Identification of Dyslexia in Children"** that is being submitted in partial fulfillment for the award of B.Tech in Computer Science Engineering to the Jawaharlal Nehru Technological University is a record of bonafide work carried out by her in our guidance and supervision.

Supervisor	CoE Incharge
Mrs. K. Sindhura Assistant Professor, Department of CSE .	Dr. A. Sharada, Professor & Head, Department of CSE.

Head of the Department
Dr. A. Sharada,
Professor, Head,
Department of CSE.

Acknowledgements

We would like to express our sincere thanks to **Dr. K. Ramesh Reddy**, Principal, GNITS, for providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. K. Rama Linga Reddy**, Professor & Dean Academics, Department of ETE, **Dr. M. Seetha,** Professor & Dean R&D, Department of CSE, **Dr. N. Kalyani,** Professor & Dean of Innovation and Incubation Department of CSE, GNITS, for all the timely support and valuable suggestions during the period of our project.

We extend our heartfelt gratitude to **Dr. A. Sharada,** Professor & Head, Department of Computer Science and Engineering, GNITS, for her unwavering support and invaluable guidance throughout our project, providing timely assistance and insightful suggestions.

We are extremely thankful to **Dr. P. Sunitha Devi, Assistant Professor** overall project coordinator, Department of CSE, GNITS for all the valuable suggestions and guidance during the period of our project.

We are extremely thankful to **Dr. P. Sunitha Devi,** Assistant Professor**, Mrs. Ch. Swathi,** Assistant Professor, Department of CSE, GNITS, project coordinators for their encouragement and support throughout the project.

We are extremely thankful and indebted to our internal guide, **Mrs. J. Srilatha, Assistant Professor, Department of CSE,** GNITS for her constant guidance, encouragement and moral support throughout the project.

Finally, we are extremely thankful to all the faculty members and staff of CSE Department who helped us directly or indirectly, parents and friends for their cooperation in completing the project work.

P.Sathwika	**(21251A0521)**
G.Manaswini	**(21251A0542)**
P.Varsha	**(21251A0565)**
K.Samhita	**(22255A0501)**

ABSTRACT

Dyslexia is a neurodevelopmental disorder affecting reading and writing skills in children, often leading to academic challenges and psychosocial difficulties if left undiagnosed. Early identification and intervention are crucial for effective management and improved outcomes. In this proposed study, aiming to leverage machine learning techniques to develop a predictive model for early detection of dyslexia in children based on cognitive assessment scores.

Evaluation of the developed models employs a variety of metrics. These metrics serve to gauge the efficacy and reliability of the machine learning algorithms in dyslexia detection. By systematically evaluating the performance of the models, the study sheds light on their ability to accurately classify dyslexic traits based on cognitive assessment data. The findings underscore the potential of machine learning techniques as valuable tools in the early identification of dyslexia, facilitating timely interventions and support mechanisms for affected children.

By harnessing the power of machine learning, this demonstrates promising avenues for the development of reliable dyslexia screening tools. Through continued refinement and validation, these approaches hold great potential in enhancing educational outcomes and quality of life for dyslexic individuals, emphasizing the importance of proactive measures in supporting their academic and personal development.

Keywords: Machine learning, Dyslexia, Neurodevelopmental Disorder, Cognitive Assessment
Domain: Machine learning (ML).

Table of Contents

Sl.No.	Topic	Page No.
	Abstract	
	List of Figures	
	List of Tables	
1.	**Introduction ……………………………………………………………**	1
	1.1 Background of the study	2
	1.2 Problem Statement	2
	1.3 Existing systems	2
	1.4 Drawbacks of Existing systems	4
	1.5 Proposed System	6
	1.6 Advantages of Proposed Systems	7
	1.7 Methodology	8
	1.8 Objectives of the project	9
	1.9 Organization of the project	9
2.	**Literature Survey...**	11
3.	**Dyslexia Screening Model...**	14
	3.1 Architecture of the System	15
4.	**Implementation..……...**	17
	4.1 Data Acquisition	17
	4.2 Data Preprocessing	18
	4.3 Model Training	19
	4.4 Model Testing	20
	4.5 Experimental Reviews	20
	4.6 Algorithm Steps	21
	4.7 Data Analysis	22
	4.8 Conclusion	23

5. Machine Learning Algorithms.. 24

5.1 Decision Tree 24

5.2 Support Vector Machine 25

5.3 Random Forest 26

5.4 Ada Boost(Adaptive Boosting) 27

5.5 Random Forest with Grid Search 29

6. Results and Discussion... 31

6.1 Classifier Performance 31
6.2 Accuracy, Precision, Recall 31
6.3 Dataset Analysis 33
6.4 Discussion 33

7. Conclusion and Future Scope.. 39

8. References.. 41

List of Figures

SI.No.	Topic	Page No.
Figure 3.1	Proposed Model Diagram	16
Figure 5.1	Graph	30
Figure 6.1	Questionnaire of Survey	34
Figure 6.2	Questionnaire of Language Vocabulary	35
Figure 6.3	Questionnaire of Memory	36
Figure 6.4	Questionnaire of Visual discrimination	36
Figure 6.5	Questionnaire of Audio discrimination	37
Figure 6.6	Dataset	38

List of Tables

SI.No.	**Topic**	**Page No.**
Table 6.1	Performance Metrics Table	32

1. INTRODUCTION

Dyslexia is a language-based learning disorder that affects an individual's ability to read, write, spell, and speak. It is part of a broader category of information-processing problems known as learning impairments, which impact learning and other cognitive functions. If not identified early, children with dyslexia may face social stigma, leading to negative self-image, frustration, and even aggression. By utilizing insights from cognitive science and computer modeling techniques, can better understand how individuals with dyslexia process information and develop supportive technologies to recognize dyslexia early.

Children with dyslexia are often misunderstood and labeled with "non-experiencing" or "compartmental issues," which can lead to frustration and low self-esteem. Despite these challenges, dyslexia does not reflect a lack of intelligence or effort. One in five children is affected by language-based dyslexia, which is primarily categorized into two forms: phonological and surface dyslexia. Phonological dyslexia involves difficulties processing the sounds of words, while surface dyslexia involves trouble recognizing whole words by sight. Early intervention significantly reduces dyslexia's impact, emphasizing the importance of understanding its causes to develop effective strategies.

Diagnosing dyslexia involves a comprehensive assessment, as it cannot be identified through a brain scan or blood test. The evaluation considers factors such as the patient's developmental, educational, and medical history, as well as the home environment and familial relationships. Surveys assessing reading and language abilities, alongside tests for vision, hearing, and neurological functions, help identify other disorders that might contribute to reading difficulties. Mental health evaluations also provide insight into emotional challenges, such as social issues and anxiety or depression, which may limit the individual's capabilities.

Understanding the causes of dyslexia is essential for developing effective remedial techniques. Dyslexia primarily stems from an impairment in phonological processing, making it challenging to establish tasks that facilitate reading. Identifying these tasks is crucial for determining which children are at risk of developing dyslexia. By advancing understanding of dyslexia through research and technology, improving early detection and intervention, ultimately supporting affected individuals in overcoming the challenges of this learning disorder.

1.1 Background of the study

Predicting dyslexia using machine learning involves utilizing advanced computational methods to identify individuals at risk for this neurodevelopmental disorder, which manifests as difficulties in reading, spelling, and phonological processing. Affecting 5-15% of the population, dyslexia is often associated with genetic and neurological factors, necessitating early detection for effective intervention. Machine learning models, particularly those employing supervised learning algorithms such as support vector machines, decision trees, and neural networks, are used to classify individuals as dyslexic or non-dyslexic based on a variety of input features. These features can include phonological awareness scores, eye-tracking data, brain imaging results, and genetic markers. By analyzing these data points, machine learning models aim to improve the accuracy and efficiency of dyslexia diagnosis, enabling personalized educational strategies and better outcomes for those affected.

1.2 Problem Statement

The early identification of dyslexia through machine learning algorithms can revolutionize the approach to educational and developmental support. By accurately diagnosing dyslexia at a young age, tailored interventions can be implemented, enhancing the individual's learning experience and reducing the impact on their academic, professional, and social development. This proactive approach not only fosters personal growth but also optimizes resource allocation in educational systems, paving the way for more inclusive and effective learning environments.

1.3 Existing systems

- **Traditional Assessment Methods**: Traditional dyslexia assessments involve a comprehensive evaluation process conducted by educational psychologists, special educators, or other trained professionals. These methods typically include the following components:

- **Screening Test**: Initial screening tests are designed to identify children who may be at risk for dyslexia. These tests often include tasks that measure phonological awareness, reading fluency, spelling, and language skills. However, screening tools can sometimes yield false positives or negatives, leading to either unnecessary concern or missed cases.

- **Standardized Tests**: These are more formal assessments that compare a child's performance to age-appropriate norms. Tests may cover reading comprehension, decoding skills, spelling, and other literacy-related abilities. Standardized tests are valuable for providing objective data but may not fully capture the nuanced difficulties experienced by individuals with dyslexia.

- **Observations and Interview**s: Teachers, parents, and the individuals themselves are often interviewed to gather information about learning history, academic performance, and any observed difficulties with reading or writing. This qualitative data is crucial for contextualizing test results and understanding the broader impact of dyslexia on daily life.

- **Cognitive and Language Assessments**: In some cases, additional tests may be conducted to assess cognitive abilities, including working memory, processing speed, and language comprehension. These assessments help differentiate dyslexia from other learning disabilities and provide a comprehensive profile of the individual's strengths and weaknesses.

Despite these limitations, traditional assessment methods remain a critical component of diagnosing dyslexia, providing valuable insights that inform educational planning and individualized interventions. Advances in technology, such as machine learning, offer the potential to enhance these methods by providing more objective and efficient screening tools.

- **Machine Learning Models in Dyslexia Prediction**: In recent years, researchers have explored the application of machine learning algorithms to predict the risk of dyslexia. These models analyze various data inputs, such as linguistic features, behavioral data, and neurological patterns, to identify individuals who are at risk of developing dyslexia. The accuracy of these models varies depending on the quality and quantity of the data, the specific algorithms used, and the criteria for dyslexia diagnosis. While some studies report promising results, achieving high accuracy remains a challenge, highlighting the need for further research and refinement of these predictive tools. Despite these challenges, machine learning offers a more objective and scalable approach to dyslexia screening, potentially allowing for earlier and more widespread identification.

1.4 Drawbacks of Existing Systems

- **Limited Data Sources:**
  - **Narrow Focus:** Many existing systems focus on a limited set of data points, such as standardized test scores or single-dimensional assessments, which may not capture the full spectrum of cognitive abilities related to dyslexia.
  - **Lack of Diverse Data:** Some systems rely on outdated or homogenous datasets that do not reflect the diversity of the population, leading to biased outcomes and reduced generalizability.

- **Ineffective Data Pre-processing:**
  - **Inadequate Data Cleaning:** Existing systems might not effectively handle missing values or noisy data, leading to inaccurate predictions and unreliable results.
  - **Poor Labeling Techniques:** Inconsistent or unclear labeling of data can confuse models, resulting in lower performance and increased error rates.

- **Overfitting Issues:**
  - **Training Data Memorization:** Some models are trained on small datasets and may simply memorize the training examples rather than learning generalizable patterns, leading to overfitting.
  - **Lack of Generalization:** Models that overfit may perform well on training data but poorly on new, unseen data, making them unreliable for real-world applications.

- **Insufficient Model Testing:**
  - **Inadequate Evaluation Metrics:** Some systems use only a few performance metrics, such as accuracy, without considering other important factors like precision, recall, and F1- score, which provide a more comprehensive evaluation

of the model’s performance.

  - **Test Set Overlap:** In some cases, test sets may inadvertently include data from the training set, leading to inflated performance metrics and a false sense of model accuracy.

- **Limited Ability for Early Detection:**

  - **Delayed Identification:** Existing systems may not be effective in identifying dyslexia early, which delays interventions and support that could improve educational outcomes.
  - **Reactive Rather Than Proactive:** Many systems only address dyslexia after it has already impacted learning, rather than providing proactive assessments and early warnings.

- **Accessibility and Usability Challenges:**

  - **Complex Interfaces:** Some systems require specialized training or equipment, making them less accessible to the general population and limiting their widespread use.
  - **Lack of Adaptability:** Existing models may not easily adapt to new data or changing conditions, limiting their long-term effectiveness and requiring frequent retraining.

- **Bias and Fairness Concerns:**

  - **Cultural and Socioeconomic Bias:** Some models are trained on datasets that do not account for cultural, linguistic, or socioeconomic diversity, leading to biased predictions that may disadvantage certain groups.
  - **Unequal Performance Across Populations:** Existing systems may perform well for certain demographics but poorly for others, raising concerns about fairness and equity in dyslexia detection.

- **High Cost and Resource Requirements:**

  - **Expensive Implementation:** Some systems require significant financial resources for implementation and maintenance, making them inaccessible for smaller institutions or underfunded programs.
  - **Resource-Intensive:** High computational requirements can limit the use of existing systems, especially in resource-constrained environments.

- **Lack of Personalization:**

  - **One-Size-Fits-All Approach:** Many existing models do not account for individual differences in learning and cognitive styles, providing generic assessments rather than personalized evaluations.
  - **Inflexibility:** A lack of customization options can result in less effective interventions tailored to individual needs.

1.5 Proposed System

The proposed system aims to utilize machine learning techniques to predict the likelihood of dyslexia in children by analyzing a combination of linguistic, cognitive, syllable audio, and visual discrimination features. The system includes:

- Data Collection: Gather a diverse dataset that includes linguistic, cognitive, syllable audio, and visual discrimination data from children, encompassing both dyslexic and non- dyslexic cases.
- Feature Extraction: Extract relevant features from the collected data, such as phonological awareness, reading fluency, syllable recognition, and visual processing metrics.
- Algorithm Selection: Employ a range of classification algorithms, including Naive Bayes, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN), to classify the likelihood of dyslexia.
- Model Development and Evaluation: Build predictive models using the selected algorithms and perform a comparative analysis to evaluate their performance. Metrics such as accuracy, precision, recall, and F1 score will be used to identify the best-performing model.
- Development of an Early Detection Interface for Dyslexia: To develop an intuitive interface for early detection of dyslexia in children, enabling timely intervention and support.

1.6Advantages of the Proposed System

- **Comprehensive Data Collection:**
  - **Variety of Data Sources:** The use of quizzes and surveys ensures a comprehensive dataset that captures multiple dimensions of cognitive skills, including language vocabulary, speed, memory, visual discrimination, and audio discrimination.
  - **Multi-Faceted Assessment:** By assessing different cognitive domains, the system provides a holistic understanding of the user's cognitive abilities, increasing the likelihood of accurately identifying dyslexia.

- **Effective Data Pre-processing:**
  - **Data Cleaning and Labeling:** The system removes unwanted data and handles missing values, which enhances the quality and reliability of the dataset. This leads to more accurate predictions by the model.
  - **Structured Labeling:** Assigning labels (0, 1, 2) based on the likelihood of dyslexia ensures that the data is effectively categorized, which is crucial for training the model to recognize patterns associated with dyslexia.

- **Robust Model Training:**
  - **Training with Labeled Data:** The model is trained using a well-labeled dataset, allowing it to learn effectively from examples and improve its ability to generalize to new, unseen data.
  - **Correlation and Adjustment:** The system uses the correlation between input and output data to iteratively adjust the model, enhancing its performance over time.

- **Accurate Model Testing:**
  - **Independent Test Set:** By using a separate test set, the model's performance is evaluated objectively, ensuring that it generalizes well beyond the training data and isn't just memorizing training examples.

- **Performance Metrics:** The use of precision, recall, and F1-score metrics provides a comprehensive evaluation of the model's performance, ensuring that it is both accurate and reliable.

- **Potential for Early Identification:**
  - **Early Intervention:** By identifying individuals with a high likelihood of dyslexia early, the system enables timely interventions, which can lead to better educational outcomes and support.
  - **Scalability:** The system can be scaled to assess large populations, making it a valuable tool for educational institutions and healthcare providers seeking to screen for dyslexia.

- **User-Friendly and Adaptable:**
  - **Adaptable System:** The model can be adapted and improved as more data becomes available, allowing it to evolve and maintain high accuracy as it encounters a wider variety of cases.
  - **Accessible:** The quiz and survey format is easy to administer, making the system accessible to a wide range of users without the need for specialized equipment or training.

- **Experimental Validation:**
  - **Empirical Support:** The experimental results demonstrate the model's effectiveness across different random state values, providing evidence of its robustness and reliability.
  - **Continuous Improvement:** By analyzing different configurations and parameters, the model can be continuously refined to achieve even better performance.

1.7 Methodology

The methodology involves several key steps:

- Data Collection: Gather a diverse dataset comprising linguistic, cognitive, audio, and visual data relevant to dyslexia. This dataset should include data from both dyslexic and non-dyslexic individuals to ensure a balanced representation.

- Data Cleaning and Preprocessing: Clean the data by handling missing values, removing noise, and normalizing and encoding features. This step ensures the data is consistent and ready for model training.

- Feature Extraction: Extract meaningful features using statistical analysis and domain knowledge. This process aims to identify patterns associated with dyslexia, enhancing the model's ability to differentiate between dyslexic and non-dyslexic individuals.

- Model Training and Optimization: Train and optimize various machine learning models, including Random Forest, Decision Tree, Support Vector Machine (SVM), and AdaBoost. Hyperparameter tuning is conducted to improve model performance.

- Model Evaluation: Evaluate the models using metrics such as accuracy, precision, recall, and F1-score. The results are compared to select the best-performing algorithm.

- Deployment: Once the best model is identified, it can be deployed for practical use, potentially assisting in the early identification of dyslexia and facilitating timely interventions.

1.8 Objectives of the project

- To gather and augment the dataset for improved model training.
- To apply and experiment with machine learning algorithms for dyslexia prediction.
- To select and train the best performing model for accurate dyslexia prediction.
- To store the prediction results (No Dyslexia, Mild Dyslexia, Dyslexia) in MongoDB for future use.

1.9 Organization of the project

The project report is organized as follows:

- **Chapter 1:** Introduction, providing the background, problem statement, existing systems, proposed system, methodology, objectives, and organization of the report.

- **Chapter 2:** Literature Review, discussing related work.
- **Chapter 3:** System Design, detailing the architecture of the proposed system.
- **Chapter 4:** Describing the methodology of the proposed system.
- **Chapter 5:** Describing the machine learning algorithms used in project.
- **Chapter 5:** Results and Discussion, presenting the findings and evaluating the performance of the system.
- **Chapter 6:** Conclusion, summarizing the project outcomes.

2. LITERATURE SURVEY

Many research works have been proposed for prediction of dyslexia using ML algorithms. In this section, survey of some the recent papers are included, highlighting on the algorithm, classifier input and evaluation results applicable.

Advance Machine Learning Methods for Dyslexia Biomarker Detection: A Review of Implementation Details and Challenges[13]

This review paper offers an in-depth analysis of recent advancements in machine learning techniques for the detection of dyslexia and its biomarkers. With the rapid evolution of deep learning methodologies in the realm of medical diagnostics, it becomes imperative to critically assess their practical implementation and the specific challenges encountered. The paper meticulously evaluates 22 selected studies, focusing on the intricacies of each method's implementation and the experimental outcomes reported. By adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol, this review ensures a rigorous and transparent approach. The PRISMA framework, which includes a four- phase flow diagram detailing the selection process, enhances the clarity and reliability of the review, allowing for a systematic exploration of the methods and their effectiveness.

The review identifies and discusses several key challenges that affect the performance and clinical applicability of state-of-the-art machine learning models in dyslexia research. These challenges encompass data quality issues, the complexity of models, the process of feature selection, and the interpretability of results. By addressing these factors, the paper aims to highlight areas for improvement and propose solutions to enhance the accuracy and relevance of machine learning models for dyslexia and its biomarkers. Through this critical examination, the review seeks to provide valuable insights that will contribute to the advancement of more effective and clinically useful diagnostic tools. The ultimate goal is to achieve higher classification performance and improve therapeutic outcomes for individuals with dyslexia, thus bridging the gap between advanced machine learning techniques and practical clinical application.

Eye-Tracking Image Encoding: Autoencoders for the Crossing of Language Boundaries in Developmental Dyslexia Detection[4]

This paper addresses the challenge of integrating different study designs by developing a novel machine learning-based pipeline and evaluating it across two distinct eye-tracking datasets. The study utilized one dataset comprising 30 subjects who read Serbian text displayed in various color configurations using a remote eye-tracking system, and another dataset consisting of 185 subjects who read Swedish text recorded with a goggle-based eye-tracking system. To bridge the gap between these disparate datasets, the data were converted into grayscale images and processed using various time window configurations to parse and visualize the signals in a 2D plane.

The developed pipeline involved training an Autoencoder neural network on images from one dataset, with the reconstruction error used to generate features for both training and testing sets. These features were then employed to train several machine learning algorithms, which were subsequently evaluated on the testing datasets. The results demonstrated a classification accuracy of 85.6% on the Serbian dataset and 82.9% on the Swedish dataset. This study's approach highlights the potential for transferring machine learning models across different eye- tracking dyslexia studies, despite variations in experimental design. The successful application of the pipeline to diverse datasets suggests that it can effectively combine and leverage data from various sources, advancing the field of dyslexia research through improved diagnostic and analytical capabilities.

Dyslexia detection in children using eye tracking data based on VGG16 network[5]

This paper presents the development of a deep convolutional neural network (CNN) designed to detect dyslexia in children aged 7–13, utilizing eye-tracking data. The study involved children reading Serbian text displayed in 13 different color configurations, including variations in background and overlay colors. The raw gaze coordinates collected during these trials were formatted into colored images, which were then used to train a deep learning model based on the VGG16 architecture. Various configurations of the CNN and trial segmentation methods were evaluated to determine the optimal setup for accurate dyslexia detection.

The method was assessed using subject-wise cross-validation, achieving a classification accuracy of 87%. The results demonstrate that combining a convolutional neural network with visual encoding of eye-tracking data yields promising outcomes for dyslexia detection, with minimal preprocessing required. This approach highlights the effectiveness of deep learning techniques in analyzing complex visual data to identify dyslexia, suggesting a significant potential for improving diagnostic accuracy and efficiency in educational and clinical settings.

Cosmic Sounds: A Game to Support Phonological Awareness Skills for Children With Dyslexia[1]

This paper presents a prospective study that significantly advances the field of special education by developing a suite of games aimed at improving phonological awareness skills in children with dyslexia. Phonological awareness is a critical component of literacy development, involving the ability to recognize and manipulate the individual sound elements within words. Children with dyslexia often experience difficulties in these areas, leading to challenges in pronunciation, spelling, and overall reading proficiency. In this study, children aged 9–12 years, along with their teacher, were actively involved as co-designers in the creation of the game "Cosmic Sounds." This participatory design process not only customized the game to meet the specific needs and preferences of its young users but also provided valuable insights into their perspectives on game design and mechanics.

The study underscores the role of Game-Based Learning (GBL) in supporting phonological awareness and enhancing literacy skills. GBL employs interactive and engaging elements to present educational content, making learning both enjoyable and effective. For children with dyslexia, GBL offers several benefits, including the use of multimodal approaches—such as visual, text-based, and auditory stimuli—that cater to diverse learning styles. The integration of storylines, rewards, clear goals, and feedback within the games helps maintain motivation and attention, addressing common challenges associated with dyslexia. Research has shown that GBL can significantly improve learning outcomes by providing a dynamic and immersive learning environment, thereby supporting the development of critical phonological skills.

3. DYSLEXIA SCREENING MODEL

The proposed model aims to assess the likelihood of dyslexia using machine learning techniques applied to data collected from quizzes and surveys. The dataset is created from responses that cover a wide range of cognitive skills, including language vocabulary, processing speed, memory, visual discrimination, and audio discrimination. By focusing on these areas, the system collects comprehensive data that provides valuable insights into an individual's cognitive profile. The use of diverse data sources ensures that the model has a rich set of information to learn from, enhancing its ability to accurately identify potential dyslexia cases.

Data pre-processing is a critical step in ensuring the model's accuracy and reliability. The process begins with cleaning the data to remove unwanted entries and handle missing values. This step is essential to avoid any adverse effects from incomplete or noisy data, which can lead to incorrect predictions. Data is then transformed into a consistent format suitable for analysis, with each entry labeled according to its likelihood of indicating dyslexia. Labels are assigned as 0 for low or no likelihood, 1 for moderate likelihood, and 2 for high likelihood. This structured labeling helps the model differentiate between varying levels of dyslexia risk, improving its predictive power.

Model training involves using the pre-processed dataset to teach the machine learning algorithm how to recognize patterns associated with dyslexia. The model is trained on input data paired with the corresponding output labels, allowing it to learn from real examples. During this process, hyperparameter tuning is employed to optimize the model's performance, ensuring that it can generalize effectively to new data. By iteratively adjusting the model based on training results, the system becomes more adept at making accurate predictions across diverse cases.

Once the model is trained, its performance is evaluated through rigorous testing on a separate dataset that was not involved in the training phase. This ensures an unbiased assessment of the model's capabilities. Performance metrics such as precision, recall, and F1- score are calculated to provide a comprehensive evaluation of how well the model performs in identifying dyslexia. By comparing predictions against true outcomes, the model's accuracy and reliability are validated, ensuring that it is well-suited for real-world applications.

The advantages of this proposed system are numerous. Firstly, the comprehensive assessment from diverse data sources allows for a more accurate and holistic understanding of an individual's cognitive abilities. Effective data pre-processing and labeling further enhance model reliability. The robust training process, combined with rigorous testing, ensures that the model is both precise and generalizable. Moreover, early detection capabilities enable timely interventions, providing significant benefits for educational and healthcare applications. Finally, the system’s scalability and adaptability make it accessible to a wide range of users, allowing for continuous improvement and refinement as more data becomes available.

3.1 Architecture of the System

The architecture of the proposed dyslexia detection model is designed to efficiently process and analyze cognitive assessment data to predict the likelihood of dyslexia. The system begins with a data ingestion layer that aggregates data from quizzes and surveys targeting various cognitive skills, including language vocabulary, processing speed, memory, visual discrimination, and audio discrimination. This layer ensures that the data is collected in a structured manner from diverse sources, providing a comprehensive dataset for analysis [3.1].

The collected raw data is then passed through a robust data pre-processing module, which is crucial for ensuring data quality. This module is responsible for cleaning the data by removing inconsistencies and handling missing values. It also performs transformations such as normalizing numerical features and encoding categorical variables, which prepares the data for effective analysis. This step ensures that the input data is consistent and free of errors, which is essential for accurate model predictions.

Once the data is pre-processed and labeled, it is fed into the core machine learning engine, which serves as the heart of the model's architecture. This engine comprises several machine learning algorithms capable of handling classification tasks, such as decision trees, random forests, and neural networks. The choice of algorithm may vary depending on the complexity and characteristics of the dataset. The model undergoes a training phase, where it learns patterns and relationships within the labeled data. During this phase, hyperparameter tuning is employed to optimize the model's performance, ensuring that it can accurately differentiate between varying degrees of dyslexia risk [3.1]. This iterative training process helps the model generalize effectively to new data, reducing the risk of overfitting and enhancing its predictive capabilities.

Additionally, data augmentation techniques are integrated into the training process to enrich the dataset and improve model robustness. These techniques may include generating synthetic data samples or introducing variations to existing data, which helps the model learn to recognize patterns under different conditions and reduces the risk of overfitting on the original dataset. The final component of the model's architecture is a validation and testing framework, which evaluates the model's performance on separate datasets that were not used during the training phase. This framework calculates key performance metrics, such as accuracy, precision, recall, and F1-score, to provide a comprehensive assessment of the model's effectiveness. By testing the model on unseen data, this framework ensures that the model's predictions are reliable and applicable to real-world scenarios [3.1]. The architecture is designed to be scalable and adaptable, allowing for continuous updates and improvements as more data becomes available. Ultimately, the model provides a powerful tool for early detection of dyslexia, enabling timely interventions that can significantly enhance educational and developmental outcomes.

.

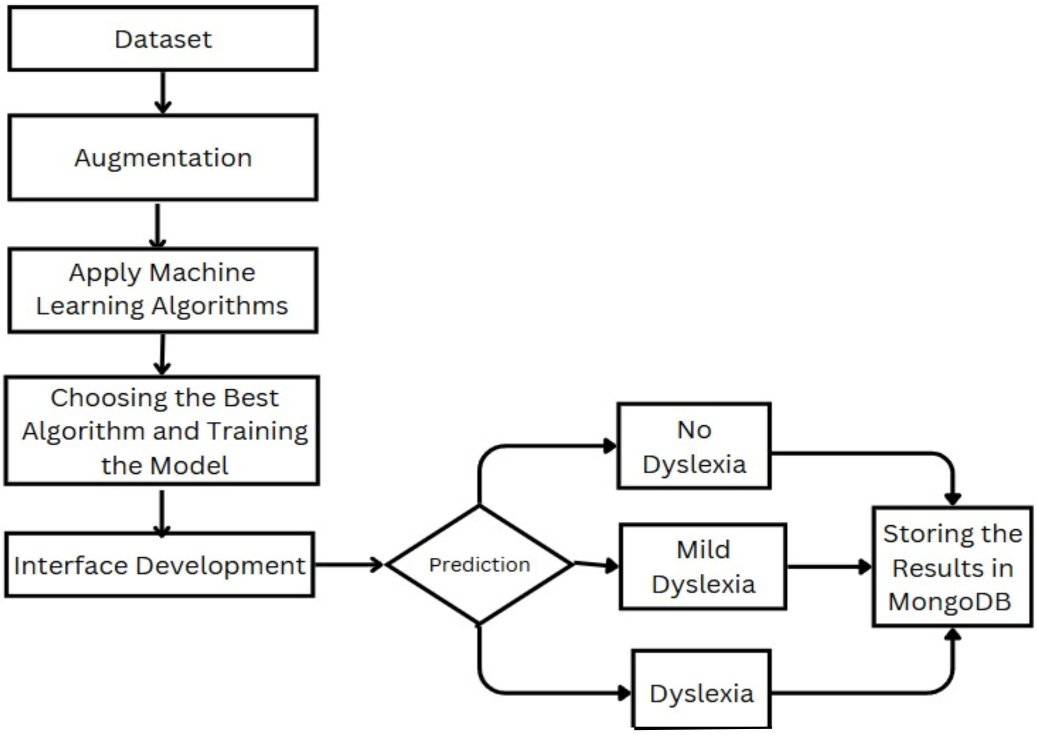

Dataset
Augmentation
Apply Machine Learning Algorithms
Choosing the Best Algorithm and Training the Model
Interface Development
Prediction
No Dyslexia
Mild Dyslexia
Dyslexia
Storing the Results in MongoDB

Fig 3.1 Proposed Model Diagram

4. IMPLEMENTATION OF MODELS

The methodology for developing a predictive tool for dyslexia using machine learning, by breaking down each stage into more detailed steps and discuss the underlying principles, techniques, and considerations involved. This expanded methodology will help ensure a systematic approach to data collection, pre-processing, model training, testing, evaluation, and results interpretation

4.1Data Acquisition:

- **Quiz and Surveys Design:**

1. **Objective Alignment:** Define clear objectives for the quizzes and surveys, targeting key dyslexia-related areas such as language processing, vocabulary, memory, visual discrimination, and auditory discrimination. Collaborate with educational psychologists and dyslexia specialists to ensure the questions are relevant and scientifically valid.
2. **Question Types:**

   Language Vocabulary: Use fill-in-the-blank, multiple-choice, or matching exercises to assess vocabulary range and understanding.

   Processing Speed: Include timed activities that measure the speed at which children can process and respond to information.

   Memory: Incorporate exercises such as recalling sequences of numbers, words, or images after short delays.

   Visual Discrimination: Design tasks that require children to identify differences in shapes, letters, or patterns.

   Audio Discrimination: Test the ability to distinguish between similar-sounding words or sounds.

   Pilot Testing: Conduct a pilot test of the quizzes and surveys with a small group of children to identify any ambiguities or issues. Gather feedback to the questions, ensuring they are age-appropriate and culturally sensitive.

- **Distribution and Data Collection**

Sampling Strategy: Use stratified sampling to ensure a diverse and representative sample of children, considering factors such as age, gender, socio-economic background, and educational setting.

Ethical Considerations: Obtain informed consent from parents or guardians and assent from the children participating in the study.Ensure anonymity and confidentiality of the participants' responses.

Administration: Decide whether quizzes and surveys will be administered online or in-person,considering accessibility and convenience. Provide clear instructions and support to participants, minimizing potential stress or misunderstanding.

- **Data Storage**

Structured Format: Store responses in a structured format I.e Excel spreadsheets .

- **Labeling**

1. **Scoring System:** Develop a comprehensive scoring system to evaluate quiz and survey responses. Assign weights to different question types based on their relevance to dyslexia indicators. Establish threshold scores for each category (Label 0, 1, and 2) using expert input and literature review.

Categorization:

Label 0: Indicates a low probability of dyslexia, suggesting typical development.

Label 1: Suggests potential dyslexic traits, warranting further observation or assessment.

Label 2: High likelihood of dyslexia, recommending a formal evaluation by specialists.

4.2Data Preprocessing:

- **Cleaning:**

- **Handling Missing Values:** Imputed missing values using the mean imputation method. This approach replaces missing data with the average value of the feature, ensuring consistency while maintaining the overall distribution of the data.

- **Scaling Techniques:** Used mean scaling to normalize the features, ensuring that the features are centered around the average value. This approach helps maintain consistency across the dataset while aligning with the characteristics of the data.

4.3Model Training:

- **Splitting the Data:**

  Training and Testing Sets: Split the dataset into 70% for training and 30% for testing using stratified sampling to preserve class distribution.

- **Training Algorithms:**

AdaBoost: Is an ensemble learning method that combines multiple weak classifiers to form a strong predictive model. It focuses on misclassified instances by adjusting their weights in each iteration, enhancing overall performance. This algorithm is particularly effective for binary classification tasks and can improve the robustness of weak learners.

Decision Tree: The Decision Tree algorithm creates a model that splits the data into branches based on feature values, resulting in a tree-like structure for decision-making. It is intuitive and easy to interpret, making it suitable for both classification and regression tasks. However, it can be prone to overfitting if not carefully tuned.

Support Vector Machine (SVM): SVM is a supervised learning algorithm that finds the optimal hyperplane to separate different classes in the dataset. It can handle both linear and non-linear classification problems by using different kernel functions, such as linear and RBF. SVM is effective in high-dimensional spaces and works well with clear margin separation.

Random Forest (Grid Search): Random Forest is an ensemble method that constructs multiple decision trees and merges their predictions to improve accuracy and reduce overfitting. It leverages the randomness in feature selection and tree construction, resulting in diverse trees. The application of Grid Search optimizes hyperparameters, enhancing the model's performance and robustness.

- **Fitting Models:**

  Model Training: Train the models using the training dataset, adjusting parameters to minimize errors and improve predictions. Monitor training progress and use early stopping techniques to prevent overfitting

4.4 Model Testing

- **Evaluation on Test Data:**

  Prediction: Use the trained models to predict the labels of the test dataset. Analyze the predictions and compare them to the actual labels.

- **Performance Metrics:**

  1. **Accuracy:** Measure the proportion of correctly predicted instances among the total instances.
  2. **Precision:** Calculate the proportion of true positive predictions among all positive predictions.
  3. **Recall:** Determine the proportion of true positive predictions among all actual positive instances.
  4. **F1-Score:** Compute the harmonic mean of precision and recall, providing a balance between them.

- **Confusion Matrix:**

  Visualization: Create a confusion matrix to visualize model performance, showing true positives, true negatives, false positives, and false negatives. Assess sensitivity (true positive rate) and specificity (true negative rate) based on the confusion matrix.

4.5 Experimental Reviews

- **Comparison of Algorithms:**

  Performance Comparison: Compare the performance of different algorithms (AdaBoost, SVM, Decision Tree, Random Forest with Grid Search) based on accuracy, precision, recall, and F1-score. Analyze the strengths and weaknesses of each algorithm concerning the specific dataset and problem domain.

- **Best Model Selection**

  Selection Criteria: Identify the model with the highest performance metrics, particularly focusing on accuracy and F1-score, to ensure robust and reliable predictions. Consider the model's interpretability, computational efficiency, and ease of deployment in real- world applications.

4.6 Algorithm Steps

1. **Label Data Set Parameters:** Set initial parameters for labeling and categorization, ensuring alignment with dyslexia indicators.
2. **Read from the Test and Survey:** Initiate values from quizzes and surveys, preparing the dataset for analysis.
3. **Split the Dataset: Divide the dataset into training and testing sets, preserving class distribution.**
4. **Apply Scaling: Normalize the data using appropriate scaling techniques.**

5. **Fit the Training Dataset:** Train the selected models on the training dataset, adjusting parameters for optimal performance.

6. **Predict Test Set Results:** Compare predicted labels to actual data labels for evaluation.

7. **Evaluate Metrics:** Assess model performance using metrics such as accuracy, precision, recall, and F1-score.

8. **Create a Confusion Matrix:** Visualize true and predicted class combinations, identifying patterns and potential areas for improvement.

9. **Assess Sensitivity and Specificity:** Evaluate sensitivity and specificity based on the confusion matrix to identify areas for improvement.

10. **Repeat Steps for All Models:** Execute steps 5-9 for all models (, Random Forest Grid Search).

11. **Model Serialization:** Save the best-performing model using techniques like joblib or pickle for future reuse and deployment.

12. **UI Integration for Prediction:** Develop and integrate a user interface (UI) that allows users to input quiz scores and instantly receive dyslexia risk predictions based on the trained model. The UI communicates with the model backend to perform real-time predictions and display results, facilitating easier accessibility for educators and parents.

4.7 Data Analysis and Results

1.AdaBoost:

Base Model: AdaBoost combines multiple weak classifiers to form a robust ensemble model. It focuses on misclassified instances by re-weighting them in each iteration, which helps improve overall model performance. This adaptive boosting approach allows the algorithm to be particularly effective in binary classification tasks.

2.Decision Tree Classifier:

Base Model: The Decision Tree algorithm splits the data into branches based on feature values, creating a tree-like structure for making decisions. This method is straightforward and easy to interpret, making it suitable for both classification and regression problems. However, it may suffer from overfitting if not properly pruned.

3.Support Vector Machines (SVM)

Support Vector Machines (SVM) are versatile classification algorithms that work well in high-dimensional spaces. They find the optimal hyperplane to separate different classes, allowing for effective classification of both linearly and non-linearly separable data.

4.Random Forest Classifier

Base Model: Random Forest builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. By averaging the results from various trees, it enhances robustness, particularly in datasets with many features

.

5.Grid Search CV: Hyperparameter tuning via Grid Search is used to optimize parameters like the number of trees, maximum depth, and minimum samples per leaf. This optimization process ensures better performance and helps the model generalize well to new data.

6.Performance Metrics

Evaluate Each Model: Each model is assessed using relevant metrics such as accuracy, recall, and F1-score. This evaluation helps identify the best-performing model for predicting dyslexia. The emphasis is on ensuring high performance across all metrics to effectively support early detection efforts.

4.8 Conclusion

This methodology provides a comprehensive approach to developing a predictive tool for dyslexia using machine learning. By systematically collecting data, pre-processing it, and applying various machine learning models, the best-performing model can be identified. This tool can aid in early diagnosis and intervention, potentially reducing the impact of dyslexia on children's academic and social development.

5. MACHINE LEARNING ALGORITHMS

5.1 Decision Trees:

A Decision Tree is a fundamental and intuitive supervised learning algorithm that is used for both classification and regression tasks. It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the input features. The tree is constructed by recursively splitting the data into subsets based on the feature that provides the best separation at each node, according to criteria such as Gini impurity, information gain, or variance reduction. Each internal node in the tree represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents a final decision or classification outcome. The primary advantage of Decision Trees is their interpretability; they allow for straightforward visualization of the decision-making process, making it easier to understand how different features influence predictions. However, Decision Trees are susceptible to overfitting, particularly with complex or noisy datasets. To counteract this, techniques such as pruning (removing parts of the tree that do not provide additional power) and setting constraints (e.g., limiting the maximum depth of the tree) are employed to improve generalization. In the context of dyslexia detection, Decision Trees are useful for identifying which specific cognitive attributes (such as vocabulary or processing speed) are most strongly associated with different levels of dyslexia risk, providing clear and actionable insights into the factors that contribute to the condition.

In the context of dyslexia detection, Decision Trees offer a powerful method for identifying key cognitive attributes that are most indicative of dyslexia risk. By analyzing various features, such as vocabulary skills, phonological processing, working memory, and processing speed, the tree can determine which attributes most strongly correlate with different levels of dyslexia risk. The interpretability of the tree allows educational and medical professionals to gain clear and actionable insights into the specific factors contributing to a child’s learning difficulties. This makes it easier to design targeted interventions and support strategies for individuals based on their unique cognitive profiles. Additionally, Decision Trees enable the visualization of the decision-making process, making it easier for professionals to understand and explain why a particular diagnosis or risk level was assigned, fostering transparency and trust in the system.

5.2 Support Vector Machine (SVM):

Support Vector Machine (SVM) is a powerful and versatile machine learning algorithm primarily used for classification tasks, but it can also handle regression and outlier detection. Its robustness makes it particularly effective when working with high-dimensional datasets and complex data structures. The underlying principle of SVM is to find the optimal hyperplane that separates data points belonging to different classes. This separation aims to maximize the margin, or the distance, between the hyperplane and the closest data points from each class. These closest points are known as support vectors, and they are critical to defining the boundary between classes. The main objective of an SVM is to find the hyperplane that maximizes the margin between classes. In the simplest case of linear separability, this hyperplane divides the feature space into two distinct regions, one for each class. The greater the margin, the better the generalization of the model to unseen data, as the model is less likely to overfit to specific training examples. Mathematically, the hyperplane is defined as:

$$w \cdot x + b = 0$$

where **w** represents the weights (or coefficients) for the features, **x** represents the feature vector, and **b** is the bias term. SVM seeks to find the values of **w** and **b** that maximize the margin between classes while minimizing classification errors. SVM is particularly effective when working with high-dimensional data, where the number of features is greater than the number of samples. In such cases, the risk of overfitting is reduced by focusing on the support vectors rather than the entire dataset. The ability to apply different kernel functions makes SVM a highly flexible algorithm, capable of handling both linear and non-linear classification tasks with ease. The choice of kernel allows SVM to capture complex relationships between features and labels.By focusing on maximizing the margin between classes, SVM is able to generalize well to new data. This margin-based approach makes it more resistant to overfitting compared to some other classifiers. In the field of dyslexia detection, SVM is highly valuable due to its ability to manage complex, high-dimensional data. Dyslexia is characterized by diverse cognitive factors, such as phonological awareness, verbal working memory, reading fluency, and visual processing, all of which can interact in non-linear ways. By using SVM with appropriate kernel functions, these intricate relationships between cognitive attributes can be modeled effectively.

SVM helps classify individuals into different risk categories (e.g., high, medium, or low risk for dyslexia) based on their cognitive profiles. By focusing on support vectors, the algorithm identifies the key cognitive features that distinguish different levels of dyslexia risk. This results in a highly accurate classification, even when the cognitive data is noisy or complex.

For example, when assessing a child's reading abilities, an SVM model can analyze the combined influence of factors like processing speed, word recognition, and phonological processing to accurately predict the child's risk of dyslexia. The flexibility of SVM, particularly with non-linear kernels like RBF, allows it to capture the subtle, non-linear patterns that might exist in such cognitive profiles.

5.3 Random Forest:

Random Forest is a powerful and widely-used ensemble learning technique that combines the strengths of multiple decision trees to enhance prediction accuracy, reduce variance, and increase the robustness of the model. By aggregating the predictions from a "forest" of decision trees, Random Forest delivers more reliable results than any individual tree could achieve. This ensemble method is highly effective for both classification and regression tasks and is known for its versatility in handling various types of data. The Random Forest algorithm operates by building numerous decision trees during the training phase and then combining their predictions. The key concept behind Random Forest is that multiple decision trees, when combined, will perform better together than individually. For classification, it takes the majority vote from the trees, outputting the class that is predicted by most of the trees (i.e., the mode).For regression, it averages the predicted values from all trees (i.e., the mean).For regression, it averages the predicted values from all trees (i.e., the mean). In the context of dyslexia detection, Random Forest can be an especially valuable tool due to its ability to manage complex, high-dimensional data involving multiple cognitive attributes. Dyslexia is influenced by various cognitive and linguistic factors, such as phonological awareness, memory, vocabulary, visual discrimination, and processing speed. These features interact in non-linear ways, making it challenging to detect patterns using simpler models.

Random Forest excels at capturing these complex relationships because it can analyze multiple features simultaneously and identify which combinations are most predictive of dyslexia risk. By utilizing the feature importance scores generated by the model, educators and clinicians can gain insights into which cognitive attributes play the most significant roles in assessing dyslexia risk. For example, if the model consistently identifies phonological processing and working memory as highly important features, interventions can be designed to specifically target these areas.

Moreover, because Random Forest does not assume linear relationships between features and outcomes, it can uncover subtle patterns in cognitive data that might be missed by other models. This makes it an effective tool for screening individuals and placing them into risk categories (e.g., high-risk or low-risk for dyslexia) based on their cognitive profiles.

Random Forest is a highly effective ensemble learning technique that combines the predictive power of multiple decision trees to deliver accurate, stable, and robust predictions. Its ability to handle high-dimensional data, reduce overfitting, and provide insights into feature importance makes it a go-to algorithm for complex classification and regression problems. In the context of dyslexia detection, Random Forest shines by analyzing intricate cognitive profiles, identifying the most influential factors contributing to dyslexia risk, and delivering accurate classification results. While it may require significant computational resources and can be less interpretable than simpler models, the advantages of Random Forest, particularly in terms of accuracy and feature analysis, make it an excellent choice for tackling complex, real-world problems.

5.4 AdaBoost (Adaptive Boosting):

AdaBoost, short for Adaptive Boosting, is a highly effective ensemble learning technique that enhances the performance of weak learners, transforming them into strong classifiers. It was developed by Yoav Freund and Robert Schapire in 1996 and is one of the most widely used boosting algorithms. The core idea behind Adaboost is to sequentially train a series of classifiers, with each new classifier focusing more on the instances that were misclassified by the previous ones. This iterative process allows the model to progressively improve its predictions, particularly for difficult or ambiguous cases. AdaBoost works by combining multiple weak learners (usually decision trees with just a few splits, also called decision stumps) to create a more powerful and accurate model. Initially, each training instance is assigned an equal weight. These weights represent the importance of each instance and are updated as the algorithm proceeds. A weak learner (a simple model that performs slightly better than random guessing) is trained on the weighted dataset. The goal of each weak learner is to minimize the classification error for the given dataset. The algorithm calculates the error rate of the weak learner, which reflects the proportion of misclassified instances. If the weak learner performs well (i.e., the error rate is low), its influence on the final prediction will be higher. If the error rate is high, the classifier's contribution will be reduced. AdaBoost increases the weights of misclassified instances so that the next classifier pays more attention to those hard-to-classify cases. By assigning more importance to incorrectly predicted data points, the algorithm encourages subsequent classifiers to focus on the most difficult instances. This helps the model correct its mistakes in future iterations. Each classifier is assigned a weight based on its accuracy (the lower the error, the higher the weight). This means that classifiers with better performance contribute more to the final prediction. The final prediction is made using a weighted sum of the predictions from all the classifiers, ensuring that more accurate models have a greater say in the outcome.After all classifiers have been trained, AdaBoost makes its final prediction by combining the outputs of all the weak learners, weighted according to their performance. For classification tasks, the final output is typically determined by majority voting, where the class with the highest weighted vote across all classifiers is selected.In the context of dyslexia detection, AdaBoost offers a powerful tool for improving the accuracy and robustness of diagnostic models. Dyslexia is a complex learning disorder that involves various cognitive factors such as reading fluency, phonological awareness, and visual processing. These factors can interact in subtle ways, making it difficult to accurately classify individuals based on their cognitive profiles. AdaBoost can be particularly valuable in this setting by focusing on challenging cases where traditional models might struggle to make accurate predictions. For instance, some individuals may exhibit atypical cognitive patterns that are harder to classify, leading to misclassifications in earlier iterations of the model. Adaboost helps correct these mistakes by refining the model to focus on these difficult cases. Dyslexia involves various interrelated cognitive factors, and AdaBoost’s sequential learning process can help capture these complexities. By combining multiple weak learners, AdaBoost can develop a more nuanced understanding of how different cognitive attributes contribute to dyslexia risk.

5.5 Random Forest with Grid Search:

Random Forest with Grid Search represents a powerful combination of machine learning techniques that leverages the Random Forest algorithm and hyperparameter tuning to optimize model performance. While Random Forest on its own is a robust ensemble learning method, the addition of Grid Search ensures that the model is fine-tuned to its full potential by systematically identifying the best hyperparameters for the task at hand. Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees to produce more accurate and stable results. It works by constructing a large number of decision trees, each trained on a different bootstrap sample of the dataset (a random subset with replacement). Each tree is also built using a random subset of the features at each node split, adding diversity to the trees and reducing the risk of overfitting. Grid Search is an exhaustive method of hyperparameter tuning that involves specifying a predefined set of hyperparameters and their potential values, then systematically evaluating the model with every possible combination of these values. The goal is to find the combination of hyperparameters that optimizes model performance.For each hyperparameter (such as n_estimators, max_depth, min_samples_leaf, etc.), a range of values is specified. For instance, you might choose to try 100, 200, and 300 trees n_estimators) and a maximum depth of 10, 20, and 30 (max_depth). The Random Forest model is trained for each possible combination of hyperparameters in the grid. This can involve a large number of models being trained if the grid contains many possible combinations. To ensure that the model generalizes well to unseen data, cross-validation is employed. Typically, the dataset is split into multiple folds, and the model is trained on some of these folds while being validated on the others. This process is repeated across all folds, and the performance metrics (like accuracy, precision, recall, and F1-score) are averaged across all folds to assess the model’s performance for each hyperparameter combination.Once all models have been trained, Grid Search evaluates the performance metrics (e.g., accuracy, precision, recall, and F1-score) to determine which combination of hyperparameters yields the best results. Based on the cross-validation results, the combination of hyperparameters that produces the highest performance is selected as the best configuration. These hyperparameters are then used to train the final Random Forest model on the full dataset.

In the context of dyslexia detection, where cognitive attributes such as reading speed, phonological awareness, and working memory are used to predict dyslexia risk, Random Forest with Grid Search provides a highly effective approach for building accurate and reliable models. Dyslexia is a complex condition that involves many interrelated cognitive factors, and identifying the most important predictors can be challenging. The Random Forest algorithm, with its inherent ability to handle complex, high-dimensional datasets, is well-suited for this task.

By using Grid Search to tune the hyperparameters of the Random Forest model, the dyslexia detection system can be optimized to achieve higher classification accuracy and better generalization to new individuals.

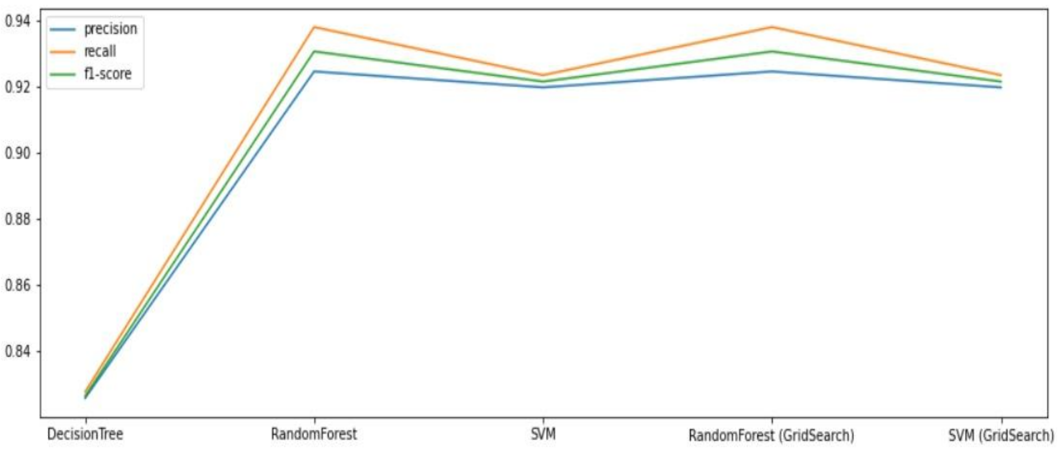


Fig 5.1 Graph

6. RESULTS AND DISCUSSIONS

The primary objective of the study was to identify the most effective machine learning classifier for predicting dyslexia risk based on demographic information and performance measures from an online gamified test. Evaluated eight popular classification algorithms: Random Forest, Logistic Regression, Extra Trees, Bernoulli Naive Bayes, K-Nearest Neighbors, Linear Support Vector Classifier, Extreme Gradient Boosting, and Light Gradient Boosting Machine.

6.1 Classifier Performance:

- Random Forest (GridSearch) consistently demonstrated the best performance across all metrics, achieving an accuracy of 92.7%, along with high precision (92.5%), recall (93.8%), and F1-score (93.1%).
- SVM Model performed similarly well, with an accuracy of 92.5%, and strong precision (92.0%), recall (92.4%), and F1-score (92.2%), closely trailing behind the Random Forest
- Decision Tree achieved moderate results with an accuracy of 83.75%, precision of 82.6%, recall of 82.8%, and an F1-score of 82.6%, but did not reach the performance level of the more complex models.
- AdaBoost yielded the lowest performance, with an accuracy of 68.25%, precision of 49.78%, recall of 47.45%, and F1-score of 45.88%, likely due to the model's limitations in handling the dataset effectively.

6.2 Accuracy, Precision, and Recall:

1. Accuracy:

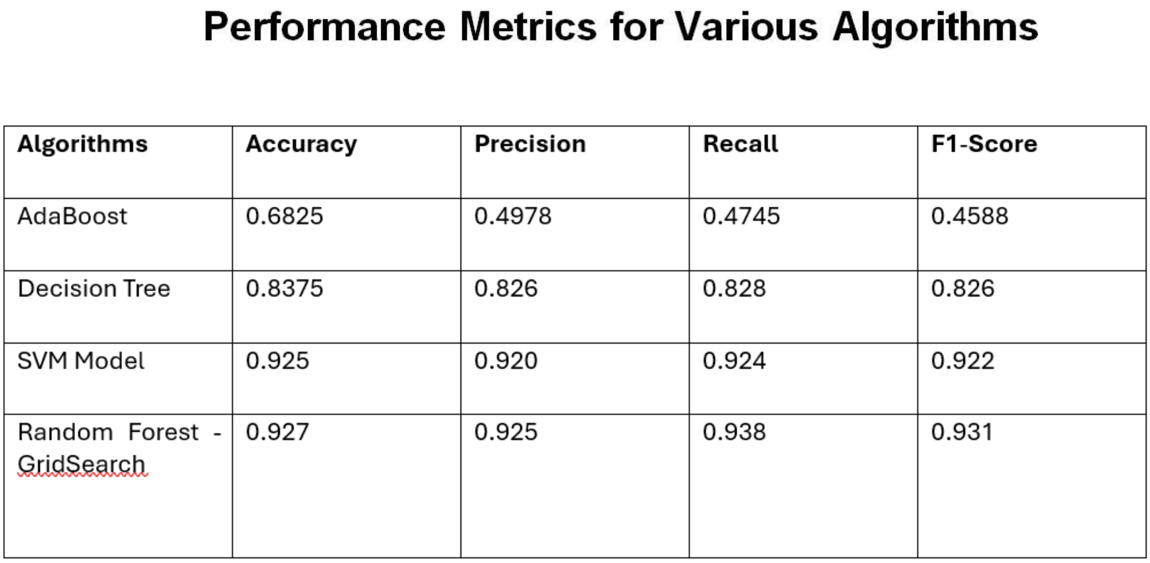
- Random Forest (GridSearch) achieved the highest accuracy of 92.7%, indicating its strong classification performance.
- SVM Model closely followed with an accuracy of 92.5%, demonstrating effective predictive capabilities.
- AdaBoost recorded the lowest accuracy at 68.25%, highlighting significant limitations in its classification ability.

2. Precision:

- The Random Forest (GridSearch) model exhibited a precision of 92.5%, showing reliability in its positive class predictions.
- The SVM Model achieved a precision of 92.0%, indicating a strong performance in accurately identifying relevant instances.
- AdaBoost had the lowest precision at 49.78%, suggesting a high rate of misclassification in its positive predictions.

3. Recall:

- Random Forest (GridSearch) led in recall with 93.8%, demonstrating its effectiveness in capturing true positive instances.
- The SVM Model achieved a recall of 92.4%, which is commendable but slightly lower than that of Random Forest.
- AdaBoost reported the lowest recall at 47.45%, indicating difficulties in identifying relevant positive cases.



Performance Metrics for Various Algorithms

Algorithms	Accuracy	Precision	Recall	F1-Score
AdaBoost	0.6825	0.4978	0.4745	0.4588
Decision Tree	0.8375	0.826	0.828	0.826
SVM Model	0.925	0.920	0.924	0.922
Random Forest - GridSearch	0.927	0.925	0.938	0.931

TABLE 6.1 Performance Metrics Table

6.3 Dataset Analysis:

- The dataset consists of 500 rows focusing on predicting dyslexia risk based on cognitive quiz scores. It includes columns for ‘Language_vocab’, ‘Memory’, ‘Speed’, ‘Visual_discrimination’, ‘Audio_Discrimination’, ‘Survey_Score’, and ‘Label’. Each applicant's scores in the first five columns reflect their abilities in language vocabulary, memory, processing speed, visual, and auditory discrimination, derived from a comprehensive quiz. The ‘Survey_Score’ is calculated from an additional survey assessing cognitive abilities. The ‘Label’ indicates the likelihood of dyslexia, with values of 0 (low), 1 (moderate), and 2 (high). This dataset provides a foundation for applying machine learning techniques to predict dyslexia risk effectively.

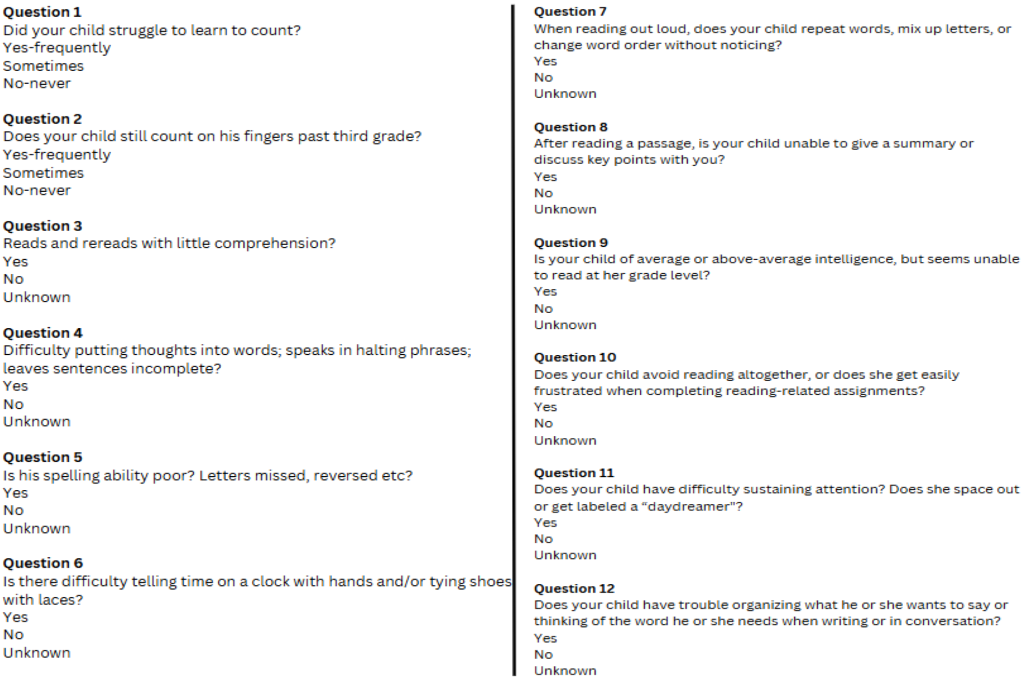
6.4 Discussion

This study focuses on the use of machine learning models to predict the likelihood of dyslexia by analyzing quiz scores and related cognitive attributes, such as language vocabulary, memory, speed, visual discrimination, and audio discrimination. The results from various models underscore several key findings and implications for dyslexia prediction.

1. **Model Performance and Selection**: Based on the performance metrics from multiple machine learning models (AdaBoost, Decision Tree, SVM, and Random Forest with Grid Search), the **Random Forest with Grid Search** model achieved the best overall performance with an accuracy of **92.7%**, a precision of **92.5%**, and a recall of **93.8%**. These metrics indicate that the model not only performed well in identifying dyslexia risk but also minimized false positives. The process of hyperparameter tuning through Grid Search played a pivotal role in optimizing the model's accuracy, making it highly suitable for practical dyslexia risk prediction.
2. **Dataset Quality and Feature Impact**: The dataset provided crucial cognitive measures that were instrumental in the success of the models. The cognitive attributes, including 'Language_vocab', 'Memory', 'Speed', 'Visual_discrimination', and 'Audio_Discrimination', allowed for a thorough assessment of dyslexia risk.

The addition of the 'Survey_Score', reflecting further cognitive evaluations, strengthened the models' predictions. This reinforces the importance of using well-structured, comprehensive datasets for accurate dyslexia screening.

3. **Importance of Feature Engineering and Tuning**: The effectiveness of feature selection and hyperparameter tuning in improving the models' performance was evident. Grid Search allowed the Random Forest model to optimize parameters such as the number of trees and depth, boosting accuracy. This process, alongside feature engineering that prioritized the most relevant cognitive attributes, was critical in producing high-quality predictions.
4. **Practical Applications and Accessibility**: One of the major strengths of this study is the creation of a user-friendly interface that facilitates real-time predictions of dyslexia risk based on quiz inputs. This tool makes it accessible for educators, parents, and non-technical users, empowering them with immediate insights into a child's dyslexia risk. However, the tool should be used as a screening aid rather than a diagnostic replacement.
5. **Questions Classification:** The certain used to classify are language, memory, speed, visual discrimination, audio discrimination, survey score. So here are the questions for each category:



Question 1
Did your child struggle to learn to count?
Yes-frequently
Sometimes
No-never

Question 2
Does your child still count on his fingers past third grade?
Yes-frequently
Sometimes
No-never

Question 3
Reads and rereads with little comprehension?
Yes
No
Unknown

Question 4
Difficulty putting thoughts into words; speaks in halting phrases; leaves sentences incomplete?
Yes
No
Unknown

Question 5
Is his spelling ability poor? Letters missed, reversed etc?
Yes
No
Unknown

Question 6
Is there difficulty telling time on a clock with hands and/or tying shoes with laces?
Yes
No
Unknown

Question 7
When reading out loud, does your child repeat words, mix up letters, or change word order without noticing?
Yes
No
Unknown

Question 8
After reading a passage, is your child unable to give a summary or discuss key points with you?
Yes
No
Unknown

Question 9
Is your child of average or above-average intelligence, but seems unable to read at her grade level?
Yes
No
Unknown

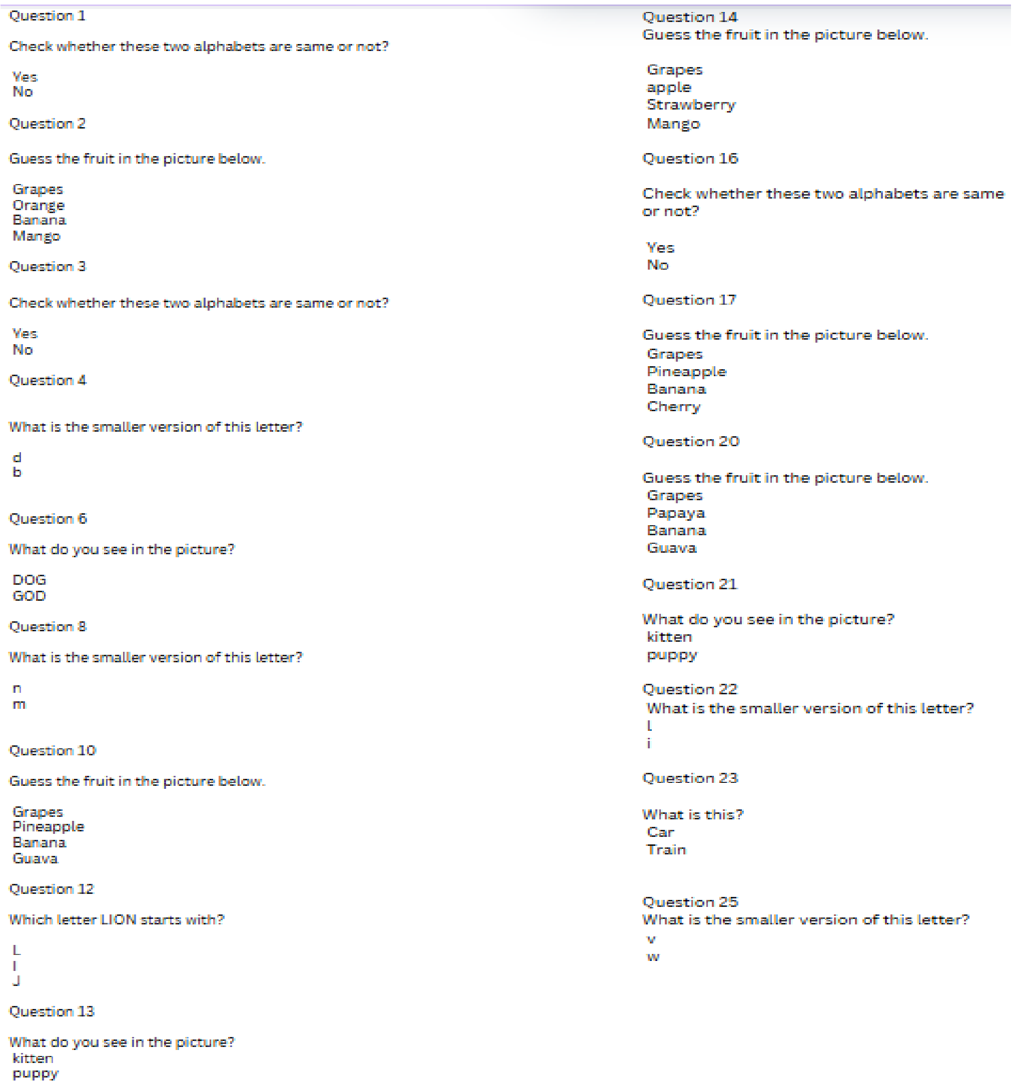
Question 10
Does your child avoid reading altogether, or does she get easily frustrated when completing reading-related assignments?
Yes
No
Unknown

Question 11
Does your child have difficulty sustaining attention? Does she space out or get labeled a "daydreamer"?
Yes
No
Unknown

Question 12
Does your child have trouble organizing what he or she wants to say or thinking of the word he or she needs when writing or in conversation?
Yes
No
Unknown

Fig 6.1 Questionnaire of Survey

The image [6.1] shows a questionnaire with 12 questions assessing learning difficulties in children, particularly around reading, counting, comprehension, and attention. Each question provides multiple-choice answers: "Yes," "No," "Sometimes," or "Unknown”.



Question 1

Check whether these two alphabets are same or not?

Yes
No

Question 2

Guess the fruit in the picture below.

Grapes
Orange
Banana
Mango

Question 3

Check whether these two alphabets are same or not?

Yes
No

Question 4

What is the smaller version of this letter?

d
b

Question 6

What do you see in the picture?

DOG
GOD

Question 8

What is the smaller version of this letter?

n
m

Question 10

Guess the fruit in the picture below.

Grapes
Pineapple
Banana
Guava

Question 12

Which letter LION starts with?

L
I
J

Question 13

What do you see in the picture?
kitten
puppy

Question 14
Guess the fruit in the picture below.

Grapes
apple
Strawberry
Mango

Question 16

Check whether these two alphabets are same or not?

Yes
No

Question 17

Guess the fruit in the picture below.
Grapes
Pineapple
Banana
Cherry

Question 20

Guess the fruit in the picture below.
Grapes
Papaya
Banana
Guava

Question 21

What do you see in the picture?
kitten
puppy

Question 22
What is the smaller version of this letter?
L
i

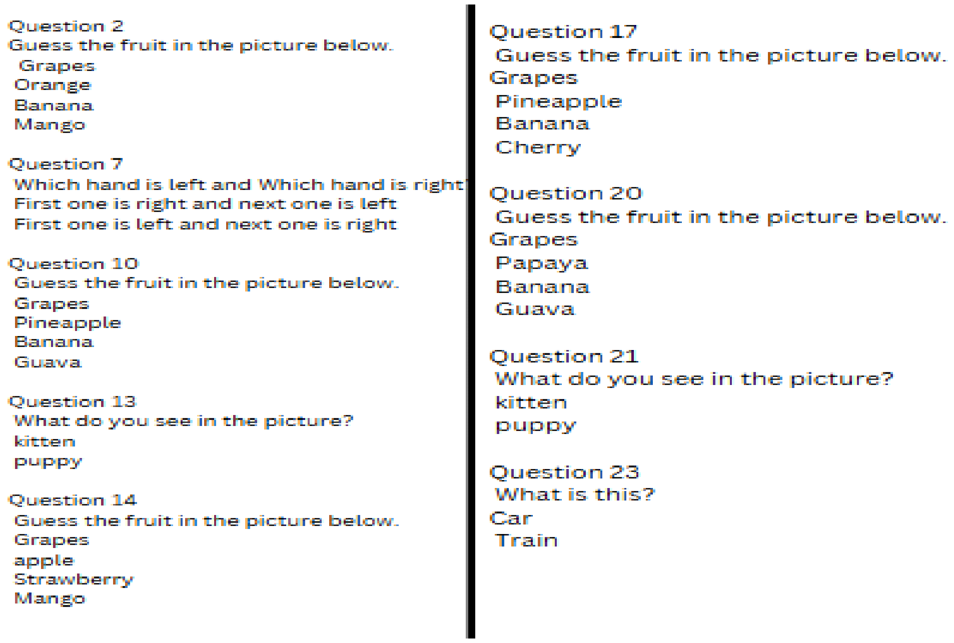
Question 23

What is this?
Car
Train

Question 25
What is the smaller version of this letter?
v
w

Fig 6.2 Questionnaire of Language Vocabulary

The image [6.2] contains a set of 25 questions focused on letter recognition, fruit identification, and comparing uppercase and lowercase letters. The questions aim to assess visual perception and basic cognitive skills in children.



Question 2
Guess the fruit in the picture below.
Grapes
Orange
Banana
Mango

Question 7
Which hand is left and Which hand is right
First one is right and next one is left
First one is left and next one is right

Question 10
Guess the fruit in the picture below.
Grapes
Pineapple
Banana
Guava

Question 13
What do you see in the picture?
kitten
puppy

Question 14
Guess the fruit in the picture below.
Grapes
apple
Strawberry
Mango

Question 17
Guess the fruit in the picture below.
Grapes
Pineapple
Banana
Cherry

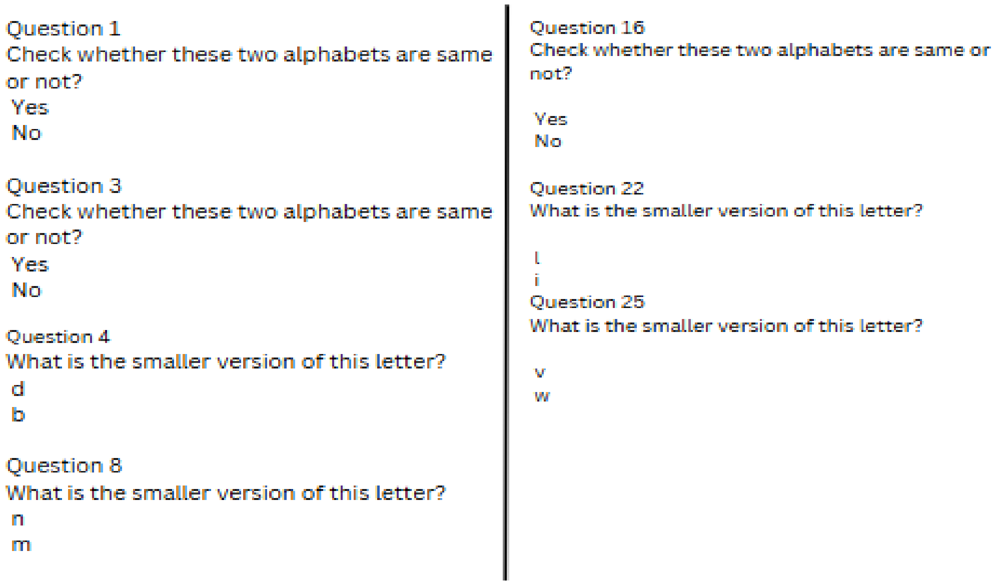
Question 20
Guess the fruit in the picture below.
Grapes
Papaya
Banana
Guava

Question 21
What do you see in the picture?
kitten
puppy

Question 23
What is this?
Car
Train

Fig 6.3 Questionnaire of Memory

The image [6.3] shows a questionnaire with various questions asking children to identify fruits, recognize left and right hands, and distinguish between animals and objects. The questions are simple and designed to assess basic cognitive and perception skills.



Question 1
Check whether these two alphabets are same or not?
Yes
No

Question 3
Check whether these two alphabets are same or not?
Yes
No

Question 4
What is the smaller version of this letter?
d
b

Question 8
What is the smaller version of this letter?
n
m

Question 16
Check whether these two alphabets are same or not?

Yes
No

Question 22
What is the smaller version of this letter?

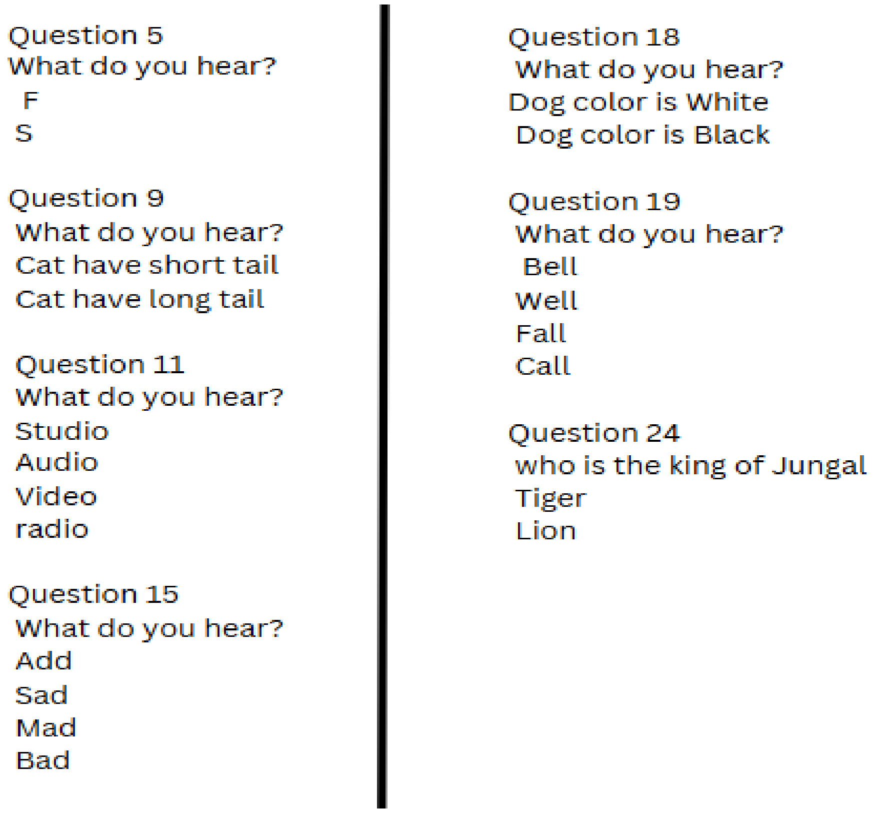
l
i

Question 25
What is the smaller version of this letter?

v
w

Fig 6.4 Questionnaire of Visual discrimination

The image [6.4] shows a questionnaire focuses on visual discrimination tasks, asking participants to identify whether two letters are the same or different and to select the lowercase version of uppercase letters. This type of activity helps assess attention to detail and letter recognition skills, which are important for reading and writing.



Question 5
What do you hear?
F
S

Question 9
What do you hear?
Cat have short tail
Cat have long tail

Question 11
What do you hear?
Studio
Audio
Video
radio

Question 15
What do you hear?
Add
Sad
Mad
Bad

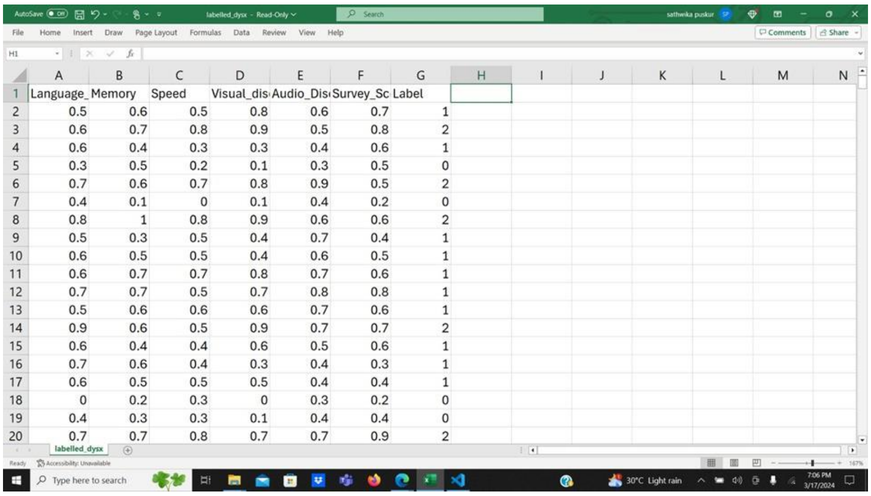
Question 18
What do you hear?
Dog color is White
Dog color is Black

Question 19
What do you hear?
Bell
Well
Fall
Call

Question 24
who is the king of Jungal
Tiger
Lion

Fig 6.5 Questionnaire of Audio discrimination

The image [6.5] shows a questionnaire focuses on audio discrimination tasks, asking participants to distinguish between sounds, such as identifying different words or phrases, and recognizing the characteristics of animal sounds. These exercises assess auditory processing and the ability to differentiate between similar-sounding words or statements.



	A	B	C	D	E	F	G
1	Language_	Memory	Speed	Visual_dis	Audio_Dis	Survey_Sc	Label
2	0.5	0.6	0.5	0.8	0.6	0.7	1
3	0.6	0.7	0.8	0.9	0.5	0.8	2
4	0.6	0.4	0.3	0.3	0.4	0.6	1
5	0.3	0.5	0.2	0.1	0.3	0.5	0
6	0.7	0.6	0.7	0.8	0.9	0.5	2
7	0.4	0.1	0	0.1	0.4	0.2	0
8	0.8	1	0.8	0.9	0.6	0.6	2
9	0.5	0.3	0.5	0.4	0.7	0.4	1
10	0.6	0.5	0.5	0.4	0.6	0.5	1
11	0.6	0.7	0.7	0.8	0.7	0.6	1
12	0.7	0.7	0.5	0.7	0.8	0.8	1
13	0.5	0.6	0.6	0.6	0.7	0.6	1
14	0.9	0.6	0.5	0.9	0.7	0.7	2
15	0.6	0.4	0.4	0.6	0.5	0.6	1
16	0.7	0.6	0.4	0.3	0.4	0.3	1
17	0.6	0.5	0.5	0.5	0.4	0.4	1
18	0	0.2	0.3	0	0.3	0.2	0
19	0.4	0.3	0.3	0.1	0.4	0.4	0
20	0.7	0.7	0.8	0.7	0.7	0.9	2

Fig 6.6 Dataset

The dataset [6.6] displayed consists of columns representing various cognitive and perceptual abilities, including Language_ Memory, Speed, Visual_ discrimination, Audio_ discrimination , and Survey_ score. The final column, Label, likely indicates classifications or groupings based on these features, with values like 1 and 2. This dataset could be used for analyzing or predicting outcomes related to these cognitive factors.

7. CONCLUSION

This study showcases the effective application of machine learning algorithms in predicting dyslexia risk using data gathered from cognitive quiz scores. Among the various classifiers examined, the Random Forest with Grid Search emerged as the most accurate and reliable model, achieving an accuracy of 92.7% with high precision (92.5%) and recall (93.8%). Its superior performance sets it apart from other models like SVM and Decision Tree, making it a highly promising tool for early detection of dyslexia.

The findings underscore the significant potential of machine learning in educational diagnostics, particularly in identifying learning disabilities like dyslexia. Using accurate and efficient models such as Random Forest with Grid Search can enhance the process of early detection, offering valuable insights into a child's cognitive abilities and learning challenges. However, while these models provide an initial screening tool, they should not replace comprehensive professional evaluations. Instead, they should be viewed as part of a broader diagnostic process, with professional assessments necessary for a definitive diagnosis of dyslexia. Additionally, the development of interactive tools based on these models could broaden the practical application of this technology in educational settings, enabling wider access to early detection tools for dyslexia.

In summary, the Random Forest with Grid Search model represents a major advancement in the use of machine learning for predicting dyslexia risk. Its accuracy and potential for real-time application offer valuable resources for educators, parents, and professionals, supporting early interventions and enabling personalized learning strategies to help children overcome learning challenges at an early stage.

Future Directions:

Future research should explore the integration of additional data sources, such as neuroimaging, to further enhance prediction accuracy. Additionally, developing interactive tools and platforms for educators and parents can facilitate early detection and intervention for dyslexia. In summary, the study demonstrates the potential of machine learning, particularly Random Forest with Grid Search, in predicting dyslexia risk with high accuracy and precision. By leveraging advanced computational techniques and comprehensive datasets, we can provide valuable support for early detection and intervention, ultimately improving outcomes for individual with dyslexia.

REFERENCES

1. A. Brennan, T. McDonagh, M. Dempsey and J. McAvoy, "Cosmic Sounds: A Game to Support Phonological Awareness Skills for Children With Dyslexia," in IEEE Transactions on Learning Technologies, vol. 15, no. 3, pp. 301-310, 1 June 2022, doi: 10.1109/TLT.2022.3170231.

2. A. S. Shany and S. Shailesh, "Learning Disability Predictions using Machine Learning: A detailed Evaluation on Predicting risk of Dyslexia," 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2023, pp. 96-101.

3. Alzamzami, F., Hoda, M., and El Saddik, A, “Light gradient boosting machine for general sentiment classification on Short texts: A comparative evaluation”, IEEE Access, vol. 8, pp. 101840–101858, May. 2020.

4. I. A. Vajs, G. S. Kvaščev, T. M. Papić and M. M. Janković, "Eye-Tracking Image Encoding: Autoencoders for the Crossing of Language Boundaries in Developmental Dyslexia Detection," in IEEE Access, vol. 11, pp. 3024-3033, 2023, doi: 10.1109/ACCESS.2023.3234438.

5. I. Vajs, V. Ković, T. Papić, A. M. Savić and M. M. Janković, "Dyslexia detection in children using eye tracking data based on VGG16 network," 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 1601-1605.

6. L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee, and Q. M. Gu, “A comparison of PCA, KPCA and Ica for dimensionality reduction in support vector machine”, Neurocomputing, vol. 55, no. 1-2, pp. 321–336, Aug. 2003.

7. L. Rello, “Predicting risk of dyslexia - PLOS One”, Kaggle, https://www.kaggle.com/datasets/luzrello/dyslexia (accessed Jan. 4, 2022).

8. L. Rello, M. Ballesteros, A. Ali, M. Serra, D. Alarcon Sánchez, and J. P. Bigham, “Dytective: Diagnosing risk of dyslexia with a game”, Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2016. doi:10.4108/eai.16-5-2016.2263338.

9. L. Rello, E. Romero, M. Rauschenberger, A. Ali, K. Williams, J. P. Bigham, and N. C. White, “Screening dyslexia for English using HCI measures and machine learning”, Proceedings of the 2018 International Conference on Digital Health, pp. 80–84, Apr. 2018.

10. L. Rello, R. Baeza-Yates, A. Ali, J. P. Bigham, and M. Serra, “Predicting risk of dyslexia with an online gamified test”, PLOS ONE, vol. 15, no. 12, Dec. 2020. doi:10.1371/journal.pone.0241687.

11. M. Rauschenberger, R. Baeza-Yates, and L. Rello, “Screening risk of dyslexia

through a web-game using language-independent content and machine learning", Proceedings of the 17th International Web for All Conference, pp. 1–12, Apr. 2020.

12. N. Ahmad, M. B. Rehman, H. M. El Hassan, I. Ahmad, and M. Rashid, "An efficient machine learning-based feature optimization model for the detection of dyslexia", Computational Intelligence and Neuroscience, vol. 2022, pp. 1–7, Jul. 2022. doi:10.1155/2022/8491753.

13. O. L. Usman, R. C. Muniyandi, K. Omar and M. Mohamad, "Advance Machine Learning Methods for Dyslexia Biomarker Detection: A Review of Implementation Details and Challenges," in IEEE Access, vol. 9, pp. 36879-36897, 2021.

14. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees", Machine Learning, vol. 63, no. 1, pp. 3–42, Mar. 2006.

15. S. Kaisar and A. Chowdhury, "Integrating oversampling and ensemble-based Machine Learning Techniques for an imbalanced dataset in dyslexia screening tests", ICT Express, vol. 8, no. 4, pp. 563–568, Dec. 2022.

16. V. Bahel, S. Pillai, and M. Malhotra, "A comparative study on various binary classification algorithms and their improved variant for optimal performance", 2020 IEEE Region 10 Symposium (TENSYMP), Nov. 2022.