School of Physics,
Engineering and
Computer Science

MSc Data Science Project
7PAM2002-0206-2023
Department of Physics, Astronomy and Mathematics

Data Science FINAL PROJECT REPORT

Project Title:

Vision for Understanding: Analysis of facial features autism detection

Student Name and SRN:

Sanshiya Rameshkumar (21066861)

Supervisor: Man Lang Tai

Date Submitted: 17th June 2024

Word Count: 11071

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

I have read the guidance to students on academic integrity, misconduct and plagiarism information at Assessment Offences and Academic Misconduct and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project **module** or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6). I have not used chatGPT, or any other generative AI tool, to write the report **or code (other than where I have** declared **or referenced it).**

I did not use human participants or undertake a survey in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: Sanshiya Rameshkumar

Student Name signature:

R. Sanshiya

Student SRN number: 21066861

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

ABSTRACT:

In an effort to improving diagnosis, the study investigates the application of machine learning models for face feature analysis-based early detection of Autism Spectrum Disorder (ASD). After evaluation of three models (XGBoost, Gradient Boosting, and Random Forest), Gradient Boosting had the highest accuracy (86.03%). This study demonstrates how machine learning can be used to speed up the diagnosis of ASD, guarantee prompt therapies, and enhance the lives of those who are affected. Incorporating these models into educational and healthcare environments can help meet the growing need for autism diagnoses and offer dependable assistance to doctors, which will ultimately strengthen the whole network of support for individuals with ASD.

The main aspect of this project, “Vision for understanding” intends to identify the facial features to streamline and enhance the identification of autism spectrum disorder (ASD). The neurological and developmental condition known as ASD has an impact on behaviour, learning, and communication. Effective intervention depends on early diagnosis, but this is frequently a drawn-out to many people in UK.

Our mission is to make sure that people with autism has to be treated in early intervention basis. "Vision for Understanding" is an approachable and trustworthy tool for early autism screening because it is intended for use in hospitals and schools. In addressing the concerns brought up by The Guardian news provider of the year, the reason behind increasing growth level of autism is a gradually developing condition right now. According to a study published in 2021, the UK saw a 787% increase in diagnoses between 1998 and 2018.

The unfortunate reality that far too many autistic people do not lead happy lives is the one thing that matters even more than the important questions surrounding autism diagnoses. People with autism have roughly 70–80% higher odds of poor mental and physical health, underachievement in school, unemployment and underemployment, victimization, social isolation, and early death than people without autism. Despite the best efforts of the NHS, the number of children waiting for an autism assessment has increased by 350% since the peak of the Covid pandemic, and waiting periods now surpass two years.

Contents

INTRODUCTION

The developmental health condition known as autism spectrum disorder (ASD) is typified by challenges with social interaction, communication, and repetitive activities. Globally, there has been an increase in the prevalence of ASD, which has raised awareness and pushed for early diagnosis and intervention. The number of people receiving diagnoses for autism in the UK has increased significantly; the NHS has seen large increases in both new and closed referrals. Early identification of ASD is essential because it allows for prompt therapies that can greatly enhance the quality of life and developmental paths of those who have the condition. Even with the growing need for diagnostic services, more effective and precise ways to detect ASD early on are still required. By using cutting-edge machine learning approaches to identify early indicators of ASD through facial feature analysis, this research seeks to close this gap.

Stage 1: General Developmental Screening:

The people with signs of ASD should visit on a frequent basis. During these checkups, general developmental screenings are performed.

Step 2: Medical Assessment

If a child has confirmed with ASD general development checks, further testing is done. Among them are:

- Neurological and medical exams
- Assessments of linguistic and cognitive abilities
- Development with the ASD caretakers assessments of everyday living abilities, including dressing, eating, and using the restroom

The initiative aims to make this procedure more efficient. By employing sophisticated machine learning and picture classification methods, can be detect early indications of ASD, even prior to the manifestation of conventional symptoms. This tool is intended for use in hospitals and schools, where it will assist in screening all children and, in the event that ASD is identified, will trigger additional medical evaluation.

By implementing our plan, this can guarantee that kids who might have ASD receive the support and care they require on time. The program can also help all youngsters because it provides a thorough developmental assessment. If our model indicates the presence of ASD, the next course of action would be a professional evaluation by a doctor. "Vision for Understanding" is a significant breakthrough in the field of early autism identification. It is a user-friendly, reliable, and efficient tool that can be used in both medical and educational settings.

According to UK NHS provided article information

- There were 143,119 patients with an open referral for autism suspected as of June 2023. 118,223 (83%) of them had a referral open for a minimum of 13 weeks.
- As of June 2023, there were 7,194 closed referrals and 10,910 new referrals. Comparing this to June 2022, there has been a 27% increase in new referrals and a 46% increase in closed referrals.
- In June 2023, 645 patients who were referred due to possible autism received an autism diagnosis, as opposed to 452 in the same month in 2022.

The immediate need to shorten the time it takes to diagnose ASD and allow for earlier intervention is mainly inspired this study. This research attempts to create an automated, non-invasive technique for identifying ASD using facial features by utilizing developments in machine learning. This method offers the potential to speed up the diagnostic procedure as well as give clinicians a reliable, impartial tool to aid in their decision-making.

Furthermore, a UK Guardian story from 2021 noted a notable increase in autism diagnoses and stressed the importance of early detection and intervention in ensuring that people with ASD receive the right care and have better long-term outcomes. The BBC also emphasized the burden that rising demand is placing on diagnostic services and the necessity for creative solutions like machine learning to ease these constraints.

The major goal of this project is to create a machine learning model that uses these facial features to distinguish between people with TD and ASD. This report describes the dataset, the analysis's methodology, and the findings, offering insights into how well HOG features identify ASD. It also discusses about how integrating all nine datasets and applying cutting-edge machine learning techniques can improve the model's performance and applicability in the future.

These tools have the potential to enhance developmental trajectories, enable prompt interventions, and ultimately improve outcomes for those impacted by the disorder. This work has the potential to have an impact outside of clinical settings by providing helpful resources for clinicians and educational settings, which will strengthen the support system for people with ASD.

REVIEW OF LITERATURE

- Kamala, K. S. Mahanaga Pooja, S. Varsha, and K. Sivapriya (2021), in their paper published in the 4th International Conference on Computing and Communications Technologies (ICCCT), DOI: 10.1109/ICCCT53315.2021.9711826 provides thorough analysis of autism spectrum disorder (ASD), a neurodevelopmental illness marked by difficulties interpreting sensory information, stereotyped behaviors, and impairments in social interaction. The writers draw attention to the difficulties that people with ASD encounter, including communication barriers, a desire for isolation, and erratic emotional reactions. They highlight the rising incidence of ASD, pointing out that the World Health Organization estimates that 1 in 160 children globally are impacted. Creating a machine learning model for accurate ASD detection is the main goal of the article in order to enable earlier intervention. The efficacy of many supervised machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes, Convolutional Neural Network (CNN), Logistic Regression, and Random Forest especially in the diagnosis of Autism Spectrum Disorder (ASD) is investigated and contrasted.

- Chistol M, Danubianu M (2021) published in (IJACSA) International Journal of Advanced Computer Science and Applications says unique method was employed to improve the early detection of Autism Spectrum Disorder (ASD) by utilizing machine learning (ML) and text mining techniques. Given the diversity of ASD's presentations and the inherent complexity and problems with delayed diagnosis, the study's goal was to give medical professionals access to cutting-edge diagnostic resources. The study's main objective was to analyze text data that worried parents sent in detailing the actions of their kids. The researchers created a dataset from a controlled experiment with 44 participants (parents of 35 boys and 9 girls with ASD diagnoses) using Rapid Miner. To find ASD patterns, they used a variety of machine learning algorithms, such as Naïve Bayes, K-Nearest Neighbors, Deep Learning, and Random Forest. Outperforming the other models, the K-Nearest Neighbors classifier got the greatest accuracy of 78.69%. This study showed how text mining, which uses parents' stories to create powerful predictive models, can be an effective, independent technique for early ASD screening.

- Mohemmed Sha M, Abdullah Alqahtani S, Shtwai Alsubai J, Ashit Kumar K(2014) published by Biomolecues explains ASD, is recognized as a complicated neurological and developmental condition that has a major influence on a person's capacity for social interaction and communication. The quality of life for people with ASD can be greatly enhanced by early intervention, but typical screening techniques including behavioral evaluations by qualified medical professionals are expensive and time-consuming. In order to categorize ASD in children and adolescents, the study suggests an advanced identification method that makes use of the Modified Bat Algorithm (MBA) in conjunction with Artificial Neural Networks (ANN), Modified ANN, Decision Trees (DT), and K-Nearest Neighbors (KNN). By automatically focusing in to identify the best solutions, the MBA increases the efficiency of the system, despite speed and accuracy issues. The suggested approach adjusts the BA optimization using random perturbation and optimal orientation in order to get beyond these restrictions. The Q-CHAT-10 dataset, which contains information for four age groups—toddlers, children, adolescents, and adults—is used to test the model. To make sure the dataset is relevant to the model, its quality is evaluated using the p-value and the Chi-Squared Statistic. According to performance measures, the updated ANN classifier outperformed other cutting-edge techniques, achieving an accuracy of 1.00. With any luck, this strong model will help doctors and researchers better diagnose ASD, which will raise the quality of life for those who with the disorder.
- Nurul Ahad Tawhid M, Siuly Siuly D, Hua Wang K, Kate Wang F, et al(2021) published by PubMed Central Recent developments in the diagnosis of Autism Spectrum Disorder (ASD) emphasize the utility of Electroencephalography (EEG) because of its accessibility, affordability, and high temporal resolution. The process of identifying autism biomarkers through traditional EEG analysis is time-consuming, subjective, and error-prone. An effective diagnostic framework using time-frequency spectrogram pictures of EEG signals was established in a study to solve these problems. Following preprocessing procedures like re-referencing, filtering, and normalizing, the raw EEG signals are converted into spectrogram images using the Short-Time Fourier

Transform. Next, both deep learning (DL) and machine learning (ML) models are used to assess these photos. Six distinct ML classifiers are used in the ML technique to classify the extracted textural features after principal component analysis has determined which ones are relevant. On the other hand, three convolutional neural network models are tested using the DL technique. Outperforming the ML-based model at 95.25% and current approaches, the DL-based model achieves a greater accuracy of 99.15%. This study indicates a viable path for computer-aided diagnosis tools by showing that DL structures may successfully identify ASD biomarkers from EEG data.

- Mahedy Hasan S, Md Palash Uddin P, Md Al Mamun T, Muhammad Imran Sharif G (2023) published by IEEE ACCESS says Four feature scaling (FS) algorithms are evaluated in the proposed framework: Quantile Transformer (QT), Power Transformer (PT), Normalizer, and Max Abs Scaler (MAS). Four ASD datasets, divided into age groups for toddlers, adolescents, children, and adults, are subjected to eight machine learning algorithms: Ada Boost (AB), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA). Using a range of evaluation parameters, the study compares performance and finds that although LDA works best for adolescents (97.12%) and adults (99.03%), AB obtains the maximum accuracy for toddlers (99.25%) and children (97.95%). Further, four Feature Selection Techniques (FSTs) are used to evaluate the feature importance: Correlation Attribute Evaluator, Info Gain, Relief F, and Gain Ratio. Based on the results, healthcare practitioners can better understand the importance of early ASD screening by carefully fine-tuning ML approaches and FS procedures to a greater extent.
- Parisa Moridian K, Navid Ghassemi P, Mahboobeh Jafari J, Salam Salloum-Asfar Tet al (2022) published by PubMed Central explains A neurological disorder known as autism spectrum disorder (ASD) affects behavior and communication in early childhood and has a variety of symptoms. Neuroimaging and psychological testing are traditional ways for detecting ASD; functional (fMRI) and structural (sMRI) magnetic resonance imaging (MRI) modalities are especially important for precise diagnosis. Even with their effectiveness, MRI-based diagnoses require a lot of work and time from

professionals. In response, a number of artificial intelligence (AI)-powered computer-aided design systems (CADS) have been created to help doctors. The two main AI methods for diagnosing ASD are conventional machine learning (ML) and deep learning (DL). This paper examines different CADS that use machine learning (ML) for automated ASD identification using MRI modalities, emphasizing the approaches and efficacy of these systems. While DL approaches have been studied extensively for ML, its application to the development of automated diagnosis models for ASD has received less attention. A summary of works using DL approaches is included in the Supplementary Appendix of the review. There are several obstacles to automated ASD diagnosis with MRI and AI, such as the requirement for big, annotated datasets, data complexity, and model correctness. A graphical comparison of DL and ML techniques shows the potential and performance variations of each method. In order to improve the early and precise diagnosis of ASD, the review ends with recommendations for future research areas that emphasize the integration of AI approaches with MRI neuroimaging. This thorough analysis highlights how AI has the potential to revolutionize the diagnosis of ASD and greatly enhance clinical practice.

- Muhammad Shoaib Farooq M, Rabia Tehseen H, Maidah Sabir S, Zabihullah Atal A corresponding (2023) published by PubMed Central represents A neurological and developmental illness that affects social and cognitive skills, autism spectrum disorder (ASD) is typified by repetitive behaviors, narrow interests, and communication challenges. In order to lessen the severity and long-term implications of ASD, early diagnosis is essential. Federated Learning (FL) has been investigated in a recent study as a potential tool for precise early ASD diagnosis. Using local data on ASD variables in both adults and children, the method comprises training logistic regression and support vector machine classifiers. The output of various classifiers is sent to a central server, where a meta-classifier ascertains the optimal technique for detecting ASD. With the help of four different ASD datasets totaling more than 600 records, the suggested FL model was able to identify ASD in adults with 81% accuracy and in children with 98% accuracy. This study emphasizes how FL may improve ASD diagnosis precision and early intervention.

- Sergio Rubio-Martín S, María Teresa García-Ordás N, Martín Bayón-Gutiérrez M, Natalia Prieto-Fernández et al (2023) published in IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) says takes a lot of resources and skilled personnel to accurately diagnose autism spectrum disorder (ASD), which is especially important for youngsters who need to receive treatment as soon as possible. For those affected, an early diagnosis can greatly enhance quality of life. This study highlights how artificial intelligence (AI) can be used to develop novel approaches to ASD diagnosis. Although conventional machine learning (ML) and deep learning (DL) methods have been used to diagnose a number of illnesses, there is still much to learn about how to utilize them to identify ASD using text analysis. Alongside AI models, such as decision trees, k-nearest neighbors (KNN), and extreme gradient boosting (XGB) for machine learning (ML) and bidirectional encoder representations from transformers (BERT) for deep learning (DL), the study makes use of natural language processing (NLP) techniques. The study entails taking tweets from Twitter users and separating texts produced by people who identify as having ASD from those who do not. A selection of 90,000 tweets for four5,000 from each group that was used for model training and testing out of a dataset consisting of 404,627 tweets. The predictive algorithms classified texts from ASD users with over 84% accuracy, according to the results. This work emphasizes how well DL models can detect and diagnose ASD, highlighting AI's critical role in developing early diagnostic techniques and eventually improving patient outcomes.

- Siva Rama Prasad K, Srinivasa Rao S, Raj Kiran K, Srinadh Reddy K et al (2023) published by 5th International Conference on Inventive Research in Computing Applications (ICIRCA) Because of its many and varied symptoms, autism spectrum disorder (ASD) is difficult to diagnose. Through the examination of behavioral and clinical data, recent developments in machine learning (ML) provide promising techniques for the diagnosis of ASD. This work investigates the application of Chi-Square feature selection to improve the K-Nearest Neighbor (KNN) algorithm for the detection of ASD. In this case, people are classified as either having ASD or not based on their attributes by the well-known classification and regression method KNN. 704 kids between the ages of 2 and 17 make up the dataset, which is split equally between kids with ASD and kids with usual development. It has twenty characteristics pertaining to conduct, communication, and social interaction. The most pertinent features for ASD identification are found using a Chi-Square feature selection method. The KNN algorithm is then trained using the chosen features, which include social smiling, gesturing, eye contact, and pointing. Using 10-fold cross-validation, the study claims an accuracy of 91.47% for the KNN method with Chi-Square feature selection. This high accuracy highlights how feature selection works to improve the performance of ML models. The study shows how machine learning techniques can be used to develop effective and dependable ways for detecting ASD. This strategy could be improved in the future by adding more features and experimenting with different ML algorithms to increase diagnostic accuracy.

- Wenjuan Wang X, Pengcheng Fu Z (2023) published by PubMed Central says The study of the human gut microbiota has attracted a lot of interest in the medical field and life sciences, especially in light of its connections to a number of human illnesses. But creating prediction models is difficult since gut microbiome relationships with conditions like autism spectrum disorder (ASD) are intricate and intertwined. Artificial intelligence (AI) provides promising techniques for processing and analyzing biologically complicated datasets, such as machine learning (ML) and deep learning. Based on the data, it was shown that the Moscow cohort obtained an AUROC value of 0.81 with 67% accuracy, while the Shenzhen cohort acquired a high AUROC value of 0.984 with 97% accuracy. The average prediction results displayed an AUROC of 0.86 and 80% accuracy when combining data from both cohorts. These results imply that a number of variables, such as dietary practices, geographic location, and population characteristics, have a substantial impact on the gut microbiota and should be taken into account when using nutritional therapy or microbial transplantation. The work highlights the potential of gut bacteria as biomarkers for diagnosing ASD and emphasizes the significance of AI in improving diagnostic approaches. The cross-cohort analysis offers insightful information on the variables influencing gut microbiota and proposes a thorough method for determining the risk of ASD. This study adds to the expanding body of research on microbiomes and how they are used in clinical ASD diagnoses.

- Manu Kohli K, Arpan Kumar K, Shuchi S (2022) published by IEEE Access DOAJ Global Trusted mentioned the detection of Autism Spectrum Disorder (ASD) has benefited greatly from technological advancements throughout the last ten years. Instead of using conventional behavioral assessments carried out by clinicians, the emphasis now is on utilizing cutting-edge technologies like deep learning (DL) and machine learning (ML) to improve the efficiency and accuracy of diagnosis. These technological advancements have made it possible to analyze enormous datasets and find minute patterns that may be signs of ASD. The evolution includes the integration of multimodal data sources, which together have increased the sensitivity and specificity of ASD diagnosis techniques. These sources include genetic data, neuroimaging, and a variety of biobehavioral markers. While data collection techniques differed throughout studies, real-world settings and controlled clinical settings were typically used to record naturalistic behaviors. In order to handle both organized (like survey data) and unstructured (like video recordings) data, processing techniques used ML and DL algorithms. The results of these research demonstrated the potential of technology-based techniques to identify at-risk infants as early as 9 to 12 months old and to diagnose ASD with high accuracy. Threats to both internal and external validity were identified, nevertheless, which emphasizes the necessity of strict validation procedures. The scoping study emphasizes how technology can revolutionize the detection of ASD. Although ML and DL approaches have demonstrated potential, a number of obstacles must be removed before they can be widely used. Strong study protocols, the implementation of cross-cultural field trials, the standardization of datasets and feature engineering techniques, and the recruitment of statistically significant participant groups are among the recommendations for further research. These actions are essential to guarantee that technology advancements in the detection of ASD are validated and extended outside of lab settings, ultimately enhancing early detection and intervention for ASD globally.
- Zhong Zhao R, Haiming Tang J, Xiaobin Zhang S, Xingda Qu Zet al (2021) published by Journal of Medical Internet Research says Prior research has demonstrated encouraging outcomes when utilizing machine learning (ML) to assess eye-tracking data in order to diagnose people with autism spectrum disorder (ASD). These

investigations usually include participants viewing still images, movies, or webpages, and they show different gaze patterns between people with ASD and normally developing (TD) people. However, gaze behavior in face-to-face conversations is very different from tasks involving the viewing of images, and there is still much to learn about the usefulness of eye-tracking data from social interactions in real life for the detection of ASD. By incorporating parameters related to ocular fixation and session length, the SVM classifier was able to reach a maximum classification accuracy of 92.31%. The use of visual fixation features alone (84.62% accuracy) or session length alone (84.62% accuracy) fared worse than this combined feature method. These findings imply that the incorporation of several behavioral characteristics improves the ability of machine learning models to differentiate between children with ASD and those without during in-person encounters. To sum up, combining ML with eye-tracking information from in-person encounters presents a viable way to improve the identification of ASD. This strategy emphasizes the potential of cutting-edge computational techniques to enhance clinical diagnosis while highlighting the significance of dynamic social environment in understanding behaviors connected to ASD.

- Song C, Zhong-Quan Jiang Z, Li-Fei Hu J, Wen-Hao Li H, et al (2022) published by PubMed Central frontiers in psychiatry says For individualized care, early identification of Autism Spectrum Disorder (ASD) with concomitant Intellectual Disability (ID) is essential, yet conventional diagnostic methods are frequently imprecise. In order to diagnose ASD with ID, this study compares machine learning (ML) techniques to conventional logistic regression (LR). 98 (40.66%) of the 241 children diagnosed with ASD also have ID. We compared four machine learning models: Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Learning. While LR had the best overall sensitivity (0.939), the SVM model had the highest accuracy (0.836) and sensitivity (0.952). Specificity-wise, SVM, RF, and XGBoost performed better than LR. All models had comparable area under the curves (AUCs), with SVM at 0.835, RF at 0.829, XGBoost at 0.845, and LR at 0.858. Decision curve analysis (DCA) revealed larger gains for both LR and SVM over a wider threshold range, with SVM demonstrating the best calibration. The results point to the potential of ML models, especially SVM, to improve early detection and intervention techniques by suggesting that they provide superior accuracy and specificity in the diagnosis of ASD with concomitant ID.

- May Alsaidi K , Nadim Obeid P , Nailah Al-Madi L , Hazem Hiary R et (2024) published by MDPI Article Eye-tracking scan paths, which provide a means of measuring eye movements quantitatively and analyzing attentional processes, have become an important diagnostic tool for ASD. The accuracy, simplicity of usage, and affordability of this approach make it a viable platform for the development of clinical biomarkers for ASD. The T-CNN-Autism Spectrum Disorder (T-CNN-ASD) deep learning model is proposed in this paper, and it uses eye-tracking scans to categorize individuals into normally developing (TD) and ASD groups. With two hidden layers and 300 and 150 neurons, respectively, the model was validated through ten cross-validation cycles with a dropout rate of 20%. With a testing accuracy of 95.59%, the T-CNN-ASD model outperformed various machine learning methods, including Multi-Layer Perceptron (MLP), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN). The findings illustrate the potential of the T-CNN-ASD model as an effective and dependable method for ASD diagnosis by showing that it can reliably and efficiently distinguish children with ASD from those with TD without the need for human interaction. The potential of combining cutting-edge machine learning algorithms with eye-tracking data to improve the precision and effectiveness of ASD screening is highlighted by this work.

- Vidiya Z, Shetty T, Dandekar K, Devnani A (2022) published by international Conference on Sustainable Computing and Data Communication Systems (ICSCDS) ASD, often known as autism spectrum disorder, is a serious developmental impairment that affects social and communication abilities. Effective intervention requires early diagnosis. Conventional diagnostic techniques frequently lack specialized medical testing and are laborious and subjective. The diagnosis of ASD may be improved by recent developments in deep learning (DL) and machine learning (ML). By examining children's attentional processes, eye-tracking technologies have produced results with high accuracy rates, up to 95.59%, when paired with deep learning models such as T-CNN-ASD. Furthermore, there have been notable increases in accuracy when ML classifiers like KNN and Chi-Square are integrated with feature selection approaches. There are still issues to be resolved, such as inconsistent data and the requirement for strong validation and clinical practice integration. up order to guarantee that ML and

DL models can deliver precise and effective ASD diagnoses, improving early intervention efforts, future research should concentrate on filling up these gaps.

- Zuqun W, Rui D, and Wang J (2021) published by MDPI Despite extensive research, the connection between vitamin D deficiency and autism spectrum disorder (ASD) is still complicated and poorly understood. The purpose of this meta-analysis was to determine quantitatively whether vitamin D levels and ASD are significantly correlated. After a thorough search of PubMed, EMBASE, Web of Science, and the Cochrane Library, 34 pertinent studies with a total of 20,580 individuals were found. With a mean difference (MD) of −7.46 ng/mL, children and adolescents with ASD had significantly lower vitamin D levels than controls, according to a meta-analysis of 24 case–control studies. Furthermore, odds ratios (OR) from ten case-control studies indicated that a higher risk of ASD was linked to decreased vitamin D levels (OR: 5.23). These results were validated by an analysis that removed data from earlier meta-analyses, showing a comparable MD of −6.2 ng/mL. Additionally, prospective studies showed a 54% higher probability of ASD development with reduced maternal or neonatal vitamin D levels. Furthermore, those who later acquired ASD also tended to have lower vitamin D levels. These results emphasize the significance of monitoring and treating vitamin D levels in ASD patients and at-risk populations, including pregnant and breastfeeding women, by highlighting the possible association between vitamin D insufficiency and elevated risk of ASD.

- Jaime Esqueda-Elizondo J, Juárez-Ramírez R, Roberto López-Bonilla O, Efrén García-Guerrero E et al (2022) published by DOAJ Open Global Trusted The neurological disorder known as autism spectrum disorder (ASD) is characterized by repetitive activities, trouble interacting with others, and communication challenges. People with ASD frequently have inconsistent attention spans because they are extremely sensitive to outside cues. The cognitive function of attention enables people to ignore unimportant stimuli and concentrate on pertinent ones. This study investigates a method for using electroencephalographic (EEG) data to measure attention in a 13-year-old kid with ASD. During a variety of learning activities, EEG data were obtained utilizing an Epoc+ Brain-Computer Interface (BCI) via the Emotiv Pro platform. Matlab 2019a was used for signal processing. In order to compute band power spectrum densities and extract features like Theta Relative Power (TRP), Alpha Relative Power (ARP), Beta Relative Power (BRP), Theta–Beta Ratio (TBR), Theta–Alpha Ratio (TAR), and Theta/(Alpha+Beta), the study concentrated on electrodes F3, F4, P7, and P8. Neurofeedback and attention detection depend on these characteristics. These features were used in the training and evaluation of several machine learning (ML) models. The model with the highest performance was the multi-layer perceptron neural network (MLP-NN), which obtained an AUC of 0.9299, a Cohen's Kappa coefficient of 0.8597, a Matthews correlation coefficient of 0.8602, and a Hamming loss of 0.0701. These findings imply that EEG-based attention evaluation can help with the creation of customized learning environments for people with ASD and offer measurable information to assist educators and therapists in tracking their progress.

- Parisa M , Navid G , Mahboobeh J , Salloum-Asfar S (2022) published by arxivlabs says The neurodevelopmental disorder known as autism spectrum disorder (ASD) is characterized by repetitive behaviors, communication difficulties, and an early onset. Neuroimaging and psychological testing are two of the many techniques used to identify ASD, with magnetic resonance imaging (MRI) being especially important. Although non-invasive, functional (fMRI) and structural (sMRI) MRIs can be labor-intensive for specialists. Several artificial intelligence (AI)-based computer-aided design systems (CADS) have been created to address these issues by combining deep learning (DL) and traditional machine learning (ML) methodologies. This paper examines the use of AI in MRI modalities for automated ASD detection. There has been little investigation of DL approaches for this purpose, despite the development of numerous CADS utilizing ML techniques. The Supplementary Appendix of the paper includes an overview of the available DL-based diagnostic models along with information on the difficulties in diagnosing ASD using MRI and AI. The efficiency of ML and DL research in automated ASD diagnosis is demonstrated through a graphical comparison. In order to improve diagnostic efficiency and accuracy, the paper makes recommendations for future research areas, highlighting the integration of AI approaches with MRI neuroimaging.

- Roberta Simeoli, Angelo Rega, Mariangela Cerasuolo, Raffaele Nappo et al (2024) published by spingerlink says the diagnosis of autism spectrum disorder (ASD) has historically been based on behavioral observations, which can be laborious and imprecise. Machine learning (ML) algorithms combined with technological screening tools have drawn interest as a way to improve the diagnostic process. Using machine learning (ML) approaches including artificial neural networks (ANN), support vector machines (SVM), a priori algorithms, and decision trees (DT), researchers have created new screening methods throughout the past 20 years. In order to build and evaluate predictive models, these techniques frequently make use of pre-existing datasets from common diagnostic instruments. Furthermore, the discovery of novel objective behavioral metrics, such as biomarkers, can enhance the efficacy of currently available screening instruments. This study highlights the potential of machine learning (ML) to enhance clinical assessment and diagnostic procedures by critically reviewing recent literature on the application of ML for motion analysis in ASD diagnosis. Results show that motion pattern ML analysis can predict ASD as accurately as conventional gold standard techniques. Even with these encouraging outcomes, using ML techniques in clinical contexts is still difficult. The review talks about these difficulties and emphasizes the necessity for more study to properly incorporate ML systems into the diagnosis of ASD at an early age.

- Bhavana Yadav K; Shreya Vishwas, Nikitha Anand, Rishab Kashyap B S et al (2023) published by 4th International Conference for Emerging Technology (INCET) says Autism Spectrum Disorder (ASD) affects the social, communication, and behavioral domains and poses serious developmental obstacles. To address these issues and offer the right solutions and support, early detection and precise diagnosis are essential. In order to improve ASD detection, recent developments have suggested combining machine learning approaches. This work suggests combining facial image analysis with behavioral features in a dual approach. The AQ-10-Child is a measure approved by the National Institute of Health Research (NHS) that combines ten behavioral aspects with eight individual attributes found to be beneficial for diagnosing ASD. The goal of this extensive feature set is to increase the precision of behavioral assessments. The pretrained CNN EfficientNet model is used to process photos of children in order to detect ASD using face recognition technology. This algorithm analyzes face traits linked to ASD by utilizing deep learning. By combining these methods, the research creates a machine learning system that is integrated and accessible through a website that is fully operational. By enabling communication between users and the administrative team, this platform may simplify the diagnostic procedure and offer an intuitive user interface for early ASD detection. According to the literature, combining face image analysis with behavioral analysis may improve diagnostic accuracy; nevertheless, for practical implementation, additional validation and user experience optimization are required.

- Melinda M, Filbert H. Juwono, I Ketut A Enriko, et al (2023) published by DOAJ Open Global Trust In the framework of Brain-Computer Interface (BCI) technology, the study focuses on identifying Electroencephalography (EEG) signals in people with Autism Spectrum Disorder (ASD) using Machine Learning (ML) algorithms. Pre-processing EEG signals to eliminate noise and artifacts, obtaining useful feature comparisons, and creating the best classification strategy are the main responsibilities. The methodology uses the Continuous Wavelet Transform (CWT), which was selected because it provides a thorough representation of EEG signals in the time-frequency domain, to decompose time-frequency signals. Two methods are used to classify EEG signals: first, CWT coefficients are used, and second, statistical parameters (mean, standard deviation, skewness, and kurtosis) generated from CWT are used. Classification is done using Support Vector Machine (SVM), a supervised learning technique that finds the best hyperplane to divide data into distinct classes. The application of CWT in conjunction with SVM produces a classification accuracy of 95%, according to the results, which is much greater than the 65% accuracy attained when utilizing CWT's statistical features alone. This work shows that ASD EEG signal classification performance can be significantly enhanced by using CWT for feature extraction and SVM for classification. The results imply that these techniques can successfully identify ASD based on brain wave features, providing a viable strategy for improving diagnostic instruments for ASD.
- Jain S Kumar H, Tripathy k et al (2024) published by International Conference on Emerging Systems and Intelligent Computing (ESIC) says Since early intervention greatly improves developmental outcomes, early identification of autism spectrum disorder (ASD) is essential in pediatric healthcare. A crucial tool for the early detection of ASD is behavioral analysis, which entails methodical observation and evaluation of a child's behavior in a variety of settings. Standardized behavioral instruments are used

by academics and medical professionals to assess and evaluate play behaviors, social communication skills, and sensory response. Behavioral analysis offers a wide-ranging perspective of social and communicative activities in natural environments, which is one of its primary benefits. It also offers ecologically sound insights about a child's functioning. This work uses a toddler screening dataset to use machine learning (ML) methods for ASD detection. According to Table 1's investigation, Support Vector Machine (SVM) performs better in detecting ASD than other machine learning methods, with Random Forest (RF) and Gradient Boosting (GB) following closely behind. SVM obtains an accuracy of 100%, followed by GB at 99% and RF at 97%, according to comparative visualization. These algorithms' balanced precision and recall levels demonstrate how useful they are for diagnosing ASD. These results imply that ML algorithms, especially SVM, provide a reliable way for early ASD screening in addition to conventional behavioral analytic techniques. By using cutting-edge machine learning techniques, ASD screening can become more accurate and efficient, which will be crucial for early intervention efforts.

- Kanchana A, Khilar R (2022) published by IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA) explains the early developmental is typified by challenges with social interaction and abnormal visual interpretation of emotional expressions. Usually, symptoms of ASD appear in the first two years of life. As of right now, there are neither effective treatments for ASD nor rapid and accurate diagnostic tests. A thorough screening tool has been created to effectively and dependably identify ASD in order to close this gap. This study uses a dataset of 704 occurrences and 21 attributes to investigate how different machine learning (ML) approaches may be applied to classify ASD in adults. The study investigates how well the algorithms J48 (DT), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), and Random Tree (RT) identify ASD. The main objective must to detect adult ASD in its early phases. The Random Forest method outperforms other machine learning techniques and delivers the best accuracy, at 99%, according to the results. As a result, the study suggests an RF-based predictive model that shows promise for high accuracy and dependability in detecting ASD in adults. The method uses AI to enhance early detection and intervention techniques, was a substantial development in the diagnosis of ASD.

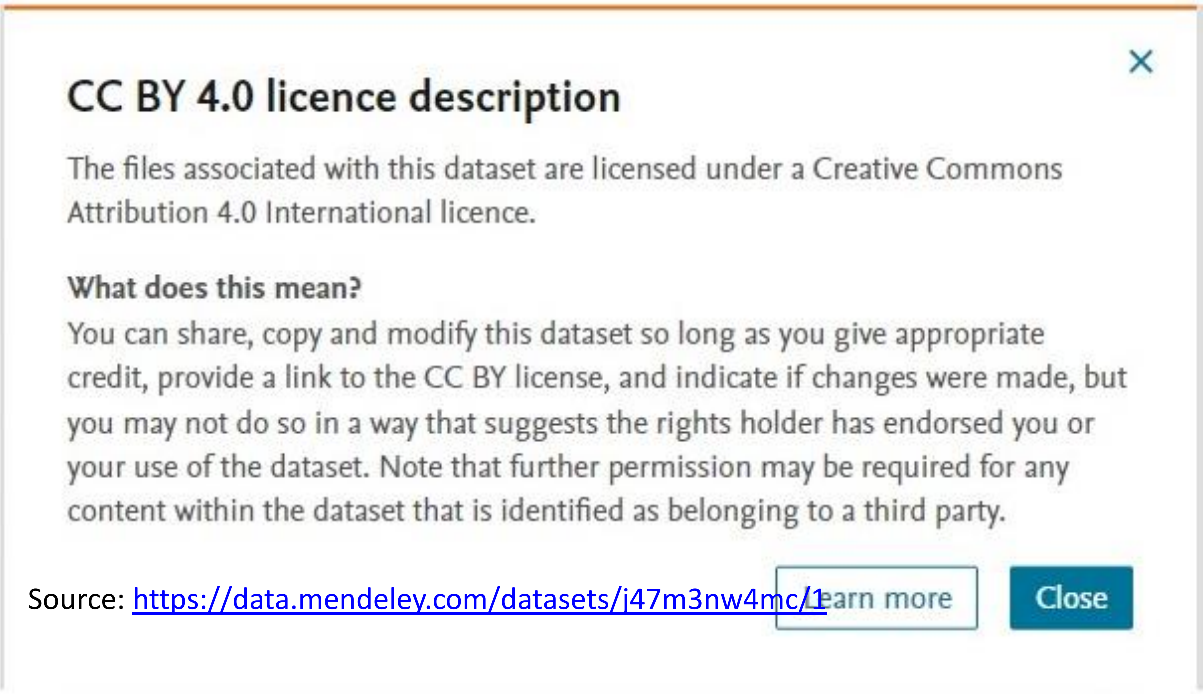
METHODOLOGY

Brief Overview:

The present research uses machine learning approaches to assess behavioral traits and facial features in an effort to improve the early detection of autism spectrum disorder (ASD). There are several important steps in the approach such as ensuring clean data for analysis, datasets containing HOG features from the eye area and entire face of individuals are first pre-processed to eliminate missing data. After that, the datasets are combined using common IDs, and to preserve data integrity, any missing values are imputed using median values. These features are used to train and assess a variety of machine learning models, such as Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and XGBoost. Metrics including classification reports, confusion matrices, and accuracy are used to evaluate the models' performance. To add more information on the models' efficacy, visualizations such as feature importance and model comparison charts are produced. After that, the top-performing model is included into a fully working web interface to enable real-time ASD diagnosis and user engagement. By combining cutting-edge technical screening methods with established behavioral assessments, this all encompass method seeks to improve the precision and effectiveness of ASD diagnosis.

Data Preparation:

The dataset contains multiple sets of facial features that were extracted using various techniques. It was obtained from the Mendeley Data repository with dataset license mentioned below.



CC BY 4.0 licence description

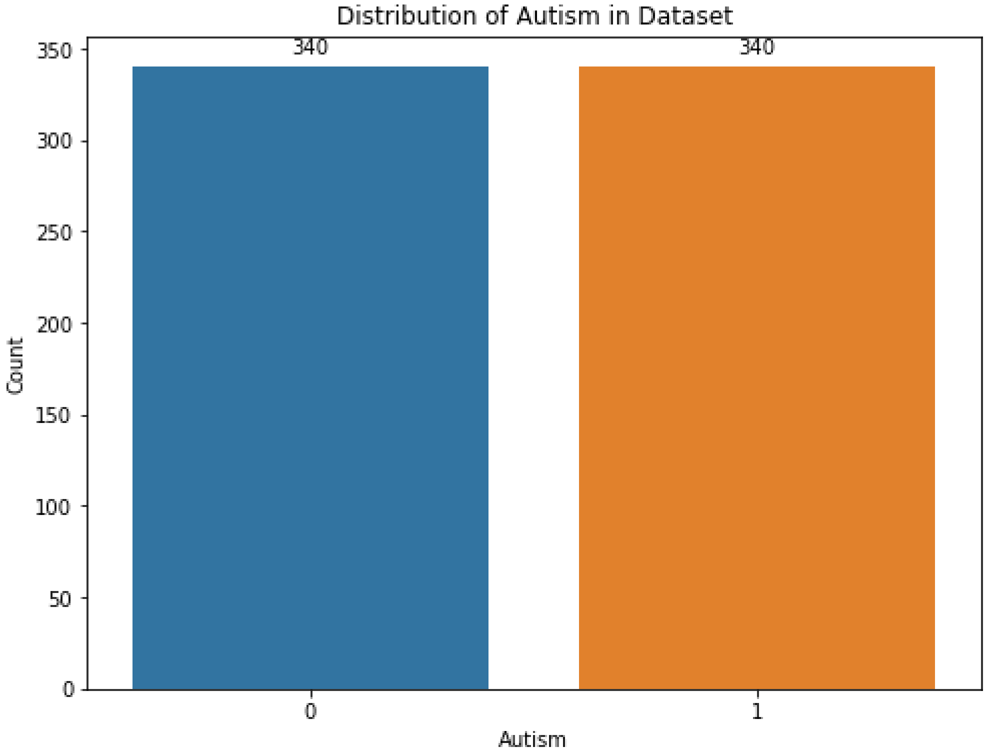
The files associated with this dataset are licensed under a Creative Commons Attribution 4.0 International licence.

What does this mean?
You can share, copy and modify this dataset so long as you give appropriate credit, provide a link to the CC BY license, and indicate if changes were made, but you may not do so in a way that suggests the rights holder has endorsed you or your use of the dataset. Note that further permission may be required for any content within the dataset that is identified as belonging to a third party.

Source: https://data.mendeley.com/datasets/j47m3nw4mc/1

The features contains Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Optical Flow (OF), Geometric Methods particularly chosen HOG eye region and the entire face, are the main subject of this report because HOG csv file are class balanced dataset which matches the motive of our autism detection to this project

There are total nine csv file which includes HOG eye region, HOG whole face, LBP eye region, LBP whole face, OF eye region, OF whole face, GEO eye region, GEO whole face and Metadata dataset where chosen HOG file path to this project total of 758 columns including index and label columns, total sample of 680 rows and most important features of HOG represents gradient intensity in different regions in the eye area.



Source: Visualised from my code

The above dataset which is visualised has binary label which determines (1) as presence and (0) as absence in the autism column which act as test data.

Ethical Consideration:

According to University of Hertforshire ethical consideration were into account as per the guidelines. All the dataset mentioned in this project are privacy and confidentiality. Furthermore, the use of machine learning models was closely supervised to protect against biases and guarantee accurate and fair outcomes across a range of demographics. The results of the study were handled with responsibility and a dedication to reporting ethics and openness. The research was conducted in accordance with the ethical guidelines established by the University of Hertfordshire's ethics committee, and any possible conflicts of interest were declared.

Data Pre-processing:

In this study, among all 9 csv file, chosen two csv file HOG whole face and eye region because it contains all the necessary features that required to accomplish the code. First, of load the appropriate dataset then renaming the first column because in both file didn’t the exact name. So, it was renamed to “id” to make it understand better. Followed by, finding common rows and column to kept them in existed scrolls for comparison and merging the two file marking as autism_x and autism_y. Next, decided to fill gaps with median value to ensure there is no missing values. Once all the gaps has been filled, removed the “id ” column where it is no longer needed then separated the attribute autism for use it in training model.

Through these steps, the data pre-processing method has successfully done and ready for training model for ASD detection.

Training the Model:

Once the dataset has been cleaned then it’s used to split and train the data. To ensure high predicted accuracy and reliability, it's essential to take three important actions when training the machine learning models. The initial step involved gathering the dataset, handling missing values through preprocessing, and encoding categorical variables. Next, the dataset was divided into testing and training sets. Three strong algorithms were chosen such as XGBoost, Gradient Boosting, and Random Forest. Using the training dataset, every model was trained to identify the underlying patterns. Each model's performance was optimized by hyperparameter tweaking, and measures including accuracy, precision, recall, and F1-score were then used to assess each model. The Gradient Boosting model was found to be the most accurate and well-balanced model based on the evaluation, which also made it the most appropriate for use in practical applications.

Splitting the Data:

The 80% of the data was utilized for training the models and the remaining 20% was used for testing. This division guarantees that the models' performance may be verified on untested data while enabling them to learn from a sizable amount of the data. Each model was trained on the training dataset using the Random Forest, Gradient Boosting, and XGBoost methods. We tweaked the hyperparameters to improve their accuracy. The test data was then used to assess the models' accuracy, precision, recall, and F1-score. The Gradient Boosting model is the most accurate and balanced performance metric achiever among them, which makes it ideal for real-time applications.

Train the Random Model Forest Model:

Random Model Forest was used at first because it suits for both of regression task and classification task. Using the strength of several decision trees, Random Forest is a potent and adaptable machine learning technique that achieves good outcomes in both classification and regression problems. It is a well-liked option in many applications because to its capacity to manage big datasets and deliver feature significance.

Train the Gradient Boosting Model:

Secondly, tried different method in contrast to bagging techniques (like Random Forest), boosting techniques construct models one after the other and concentrate on fixing mistakes from earlier models. Its application of gradient descent optimization minimizes the loss function, rendering it a potent method for error reduction. A strong and adaptable machine learning method called gradient boosting creates models one after the other to fix mistakes in earlier models. It is popularly utilized for a variety of predictive modeling applications due to its excellent accuracy. For best results, regularization and hyper-parameter adjustment must be done carefully to avoid overfitting.

Train the XGBoost Model:

Thirdly, still would like to improve the model to understand the model better chosen Extreme Gradient Boosting, or XGBoost, is a distributed gradient boosting library that has been tuned for maximum efficiency, versatility, and portability. In order to increase accuracy and performance, it incorporates a number of sophisticated features and optimizations while still building on the ideas of gradient boosting. Large datasets and complicated models are not perfect match for XGBoost, a strong and effective gradient boosting implementation. Because of its great accuracy and adaptability, it is a well-liked option for both practical applications and machine learning competitions. To fully utilize it, proper hyperparameter tuning and an understanding of regularization techniques are essential.

Evaluating the Model:

A machine learning model's performance on a particular dataset is evaluated. Finding out how effectively the model generalizes to fresh, untested data is the aim. Accuracy, precision, recall, and F1 score are often employed metrics for assessment, especially in classification tasks.

Precision:

- The ratio of accurately predicted positive observations to all expected positives is known as precision.

$$\text{Precision} = \frac{TP}{TP + FP}$$

-
- Low false positive rates are indicative of high precision. It's critical in situations (like spam detection) where the cost of a false positive is substantial.
- It can better appreciate the accuracy of the spam predictions with precision.

Recall:

- The ratio of accurately predicted positive observations to all observations made during the actual class is known as recall.

$$\text{Recall} = \frac{TP}{TP + FN}$$

-
- Low false negatives are indicative of high recall. It matters in situations when it would be expensive to overlook a positive case (such as detection).accurately predicted that every observation made in the real class would be positive.
- Recall indicates how successfully every spam email is captured by the model.

F1-Score:

- The harmonic mean of recall and precision is the F1 score. It offers a solitary metric that harmonizes the two issues**.**

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

-
- When weighing the trade-off between recall and precision, the F1 score comes in handy, particularly in cases when the distribution of classes is not uniform.

Classification Report:

1. Random Forest Model Accuracy:

86% percent of the time, the model accurately predicts whether an event is good or negative. This indicates that roughly 86 of every 100 predictions are accurate.

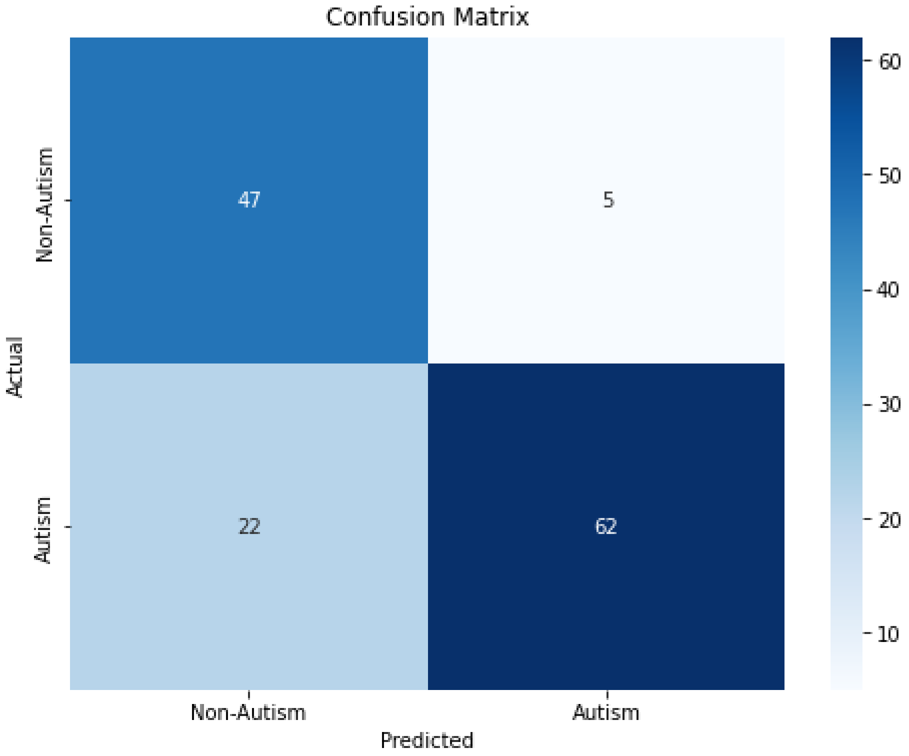
Class	Precision	Recall	F1-Score	Support
0	0.8	0.90	0.78	52
1	0.93	0.74	0.82	84
Accuracy			0.80	136
Macro Avg	0.80	0.82	0.80	136
Weighted Avg	0.83	0.80	0.80	136

Class 0 (Negative Instances):

- It is 68% precision. (The model is 68% accurate when it predicts a negative.)
- 90% recall (The model detects 90% of all real negatives correctly.)
- 78% is the F1-Score. (A harmony of recollection and precision.)

Class 1 (Positive Instances):

- There is 93% precision. (The model is 93% accurate when predicting a positive.)
- 74% of recall (74 percent of all actual positives are properly identified by the model.)
- 82% is the F1-Score. (A harmony of recollection and precision.)



Source: Confusion Matrix from

- The model properly recognized 47 negative cases, which we refer to as True Negatives (TN).
- Five negative cases were mistakenly classified as positive by the False Positives (FP) model.
- 22 positive examples were mistakenly classified as negative by the False Negatives (FN) model.
- The 62 positive examples were accurately detected by the True Positives (TP) model.

Clarify it simply, the model does exceptionally well at accurately detecting positive cases (high precision), but it might do better at detecting all positive cases (greater recall). It correctly diagnoses the majority of negative cases, while occasionally it mislabeled certain negatives as positives. The model works well overall, but it may be better, especially in terms of recall for positive cases and precision for negative ones. Though, there is potential for improvement in terms of accurately identifying all positive cases without wrongly classifying them as negative, the model generally produces trustworthy predictions.

2. **Gradient Boosting Accuracy:**

With an accuracy of 86%, the Gradient Boosting model accurately predicted the result in 86 out of 100 situations.

Classes	Precision	Recall	F1-Score	Support
0	0.77	0.90	0.83	52
1	0.93	0.83	0.88	84
Accuracy			0.86	136
Macro Avg	0.85	0.87	0.86	136
Weighted Avg	0.87	0.86	0.86	136

- With a precision of 77%, accurate when predicting a negative outcome.
- 90% of recall means real negative cases are accurately identified by the model.
- An 83% F1-Score indicates an excellent performance in identifying negative instances, balancing precision and recall.

Classes 0 (Negative Instances):

- With precision of 77%, recall is 90%, F1-score is 83%

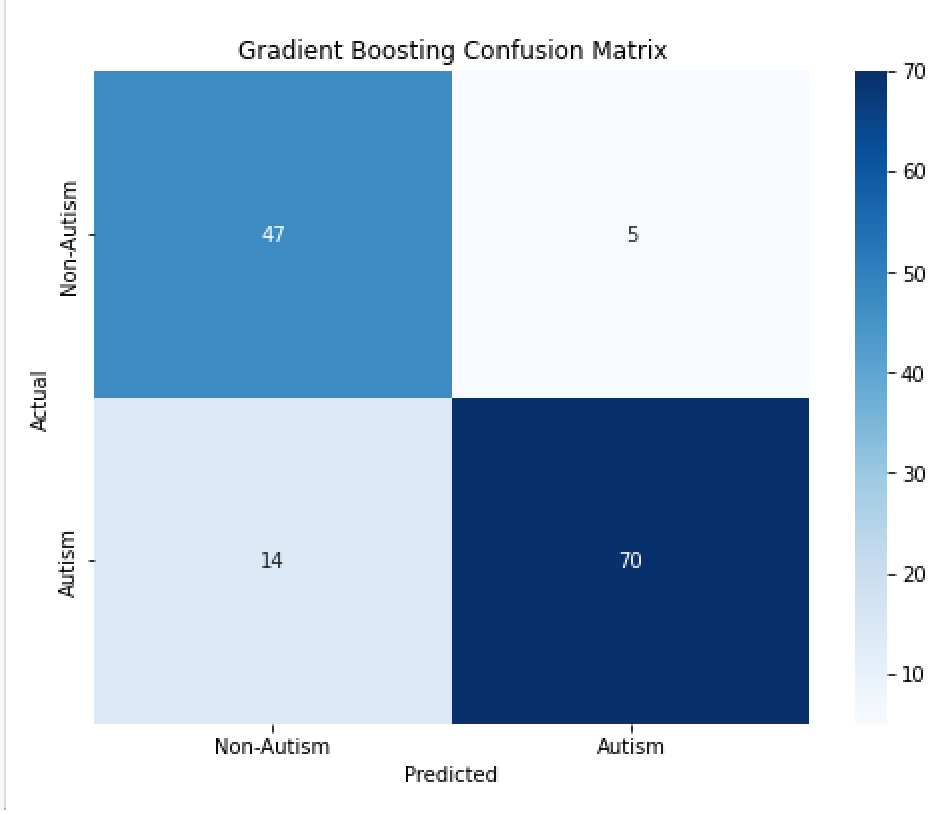
Classes 1 (Positive Instances):

- With precision of 93%, recall 83%, F1-Score is 88%

Macro Avg precision is 85%, recall is 87%, f1-score is 86% whereas Weighted Avg Precision is 87%, recall is 86%, f1-score is 86%

Overall, the model's performance is excellent, with an accuracy of 86%. With a high precision of 93%, it is very good at predicting positive examples (Class 1) and produces extremely few false positive mistakes. At 90%, the recall for negative instances (Class 0) is strong, suggesting that the model performs well in detecting the majority of true negative occurrences. Both classes have excellent F1-Scores, indicating a good recall and precision ratio.

In conclusion, the Gradient Boosting model is a dependable model for the provided dataset since it is good at producing accurate predictions, with high precision for positive cases and high recall for negative ones.



Source: From code

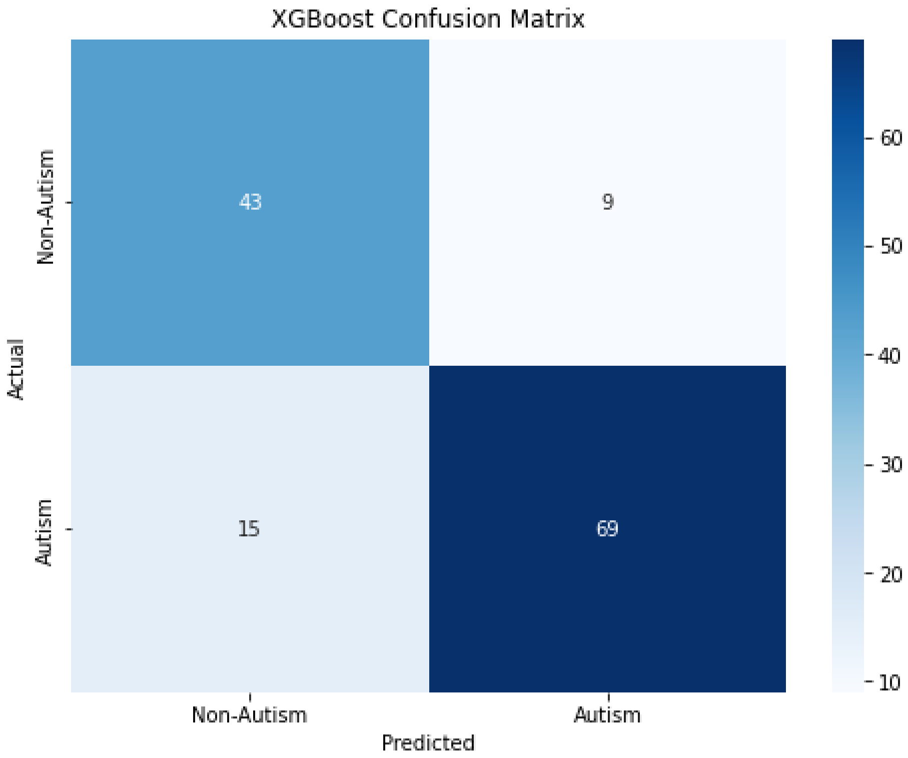
The Gradient Boosting model's confusion matrix indicates that its overall accuracy is 86%. Out of 84 positive examples (class 1), it accurately detected 70, and 47 out of 52 negative instances (class 0). As a result, the model has a high precision of 93% for positive predictions, which means that 93% of the time it is accurate when it makes a positive forecast. The algorithm properly detects 83% of all real positives, according to the 83% recall for positive predictions. The model performs well at properly recognizing the majority of negative cases for negative predictions, with a precision of 77% and recall of 90%; but, it occasionally labels positives incorrectly as negatives. Overall, the model performs well in recognizing both classes, striking a good mix between recall and precision.

3. XGBoost Accuracy

With an accuracy rate of 82.35%, the XGBoost model can accurately predict the result in about 82 out of every 100 cases.

Classes	Precision	Recall	F1-Score	Support
0	0.74	0.83	0.78	52
1	0.88	0.82	0.85	84
			0.82	136
Macro Avg	0.81	0.82	0.82	136
Weighted Avg	0.83	0.82	0.83	136

The classification report provides an overview of the XGBoost model's performance, showing that it obtained an overall accuracy of 82.35%. Its accuracy for negative cases (Class 0) is 74%, recall is 83%, and F1-score is 78%. Its greater accuracy of 88%, recall of 82%, and F1-score of 85% are for positive examples (Class 1). The model performs well overall in both classes, as evidenced by the weighted averages, which are slightly higher, and the macro average precision, recall, and F1-score, which are all between 81 and 82%.



Source: Confusion Matrix from code

Based on the XGBoost model's confusion matrix, 43 out of 52 negative occurrences (Class 0) and 69 out of 84 positive examples (Class 1) were properly detected. This yields an accuracy of 82.35% overall. For positive predictions, the model has a high precision of 88%, which means that 88% of the time it is correct when predicting a positive event. The model accurately detects 82% of all real positives, according to the 82% recall for positive events. In terms of negative cases, the model delivers an 83% recall rate, accurately identifying 83% of the actual negatives, and a 74% precision rate, indicating that 74% of the forecast negatives are accurate. With precision and recall balanced at 78% for negatives and 85% for positives, the model's F1 scores show that it can identify both classes well, with a particularly good performance in positive prediction. The XGBoost model is generally dependable and efficient, while there is potential for improvement in terms of lowering false positives and false negatives.

Limitations:

This study has a number of drawbacks even though it shows how machine learning techniques may be used to identify behavioral behaviors and facial features associated with autism spectrum disorder (ASD). The size and quality of the dataset are the main factors to consider because a little and perhaps biased dataset could make it more difficult for the model to generalize across different populations. Even with a rigorous approach, the feature selection process may miss important characteristics, and depending just on static features like Optical Flow (OF) that may fail to identify the dynamic behavioral subtleties that are crucial for the identification of ASD. Moreover, interpretability issues with the models especially with complicated models like Random Forest and XGBoost which make it hard for physicians to comprehend the decision-making process. Validating the models across many independent datasets from different demographics is vital to ensure their generalizability, and real-world application tests are required to evaluate their robustness and practical usefulness. To give a more comprehensive picture and boost diagnostic precision, the project might also profit from the integration of multimodal data sources, such as genetic, neuroimaging, and physiological data.

From a coding perspective, there is potential for improvement in the data pre-treatment method, notably with regard to handling outliers and missing values. The existing approach may not always be the best choice because it uses medians to fill in missing numbers. A more thorough study could produce more insightful results because feature importance analysis is largely restricted to the top 10 features. Further investigation into a wider range of methods and more comprehensive hyperparameter tuning may improve the models' performance. Although useful, visualizations could be improved by including more intricate depictions like ROC curves and SHAP values for easier interpretation in future.

Future Scope:

The current work was mainly concentrated on HOG dataset to find autism detection where LBP file path is good for local texture patterns such can combine with HOG file to analyse the shape or texture of an image. Likewise, OF csv file helps to measure the motion of facial characteristics especially between the frames in a video. In a same way, GEO csv file extract distances and angles between face. Each dataset represents different goals and enhance the model accuracy and robustness by different feature extraction method. Mainly these datasets can be beneficial for clinical diagnosis, educational research and social robotics.

- In addition to face feature analysis, a number of additional AI techniques have been investigated and seem promising for the identification and diagnosis of autism spectrum disorder (ASD). Natural Language Processing (NLP), for example Text and Speech Analysis in NLP approaches can be used to identify signs of ASD by analyzing speech patterns, language use, and communication abilities from written text or audio recordings. Tools that assess sentence structure, prosody, word choice, and conversational dynamics
- Brainwave analysis using electroencephalography (EEG): Artificial intelligence (AI) models are able to analyze EEG signals and recognize abnormal brainwave patterns linked to ASD. Wavelet transform and machine learning classifiers (e.g., SVM, deep learning) are two techniques that have been used to differentiate between ordinarily developing people and those with ASD.
- Social robots, such as AI-powered robots, can participate in therapeutic activities with kids who have ASD and collect data on their social interactions and reactions. This data can then be processed to track and evaluate the kids' improvement.
- Behavior Analysis in Activity Recognition: Machine learning models are able to identify and categorize patterns of behavior from motion sensors or video recordings in order to identify social interactions, repetitive activities, and other characteristics associated with ASD.
- Wearable Sensors in Physiological Monitoring: Artificial intelligence (AI) is able to identify stress and anxiety, which are frequently higher in people with ASD, by analyzing data from wearable sensors that monitor physiological responses such skin conductance, heart rate variability, and movement patterns.

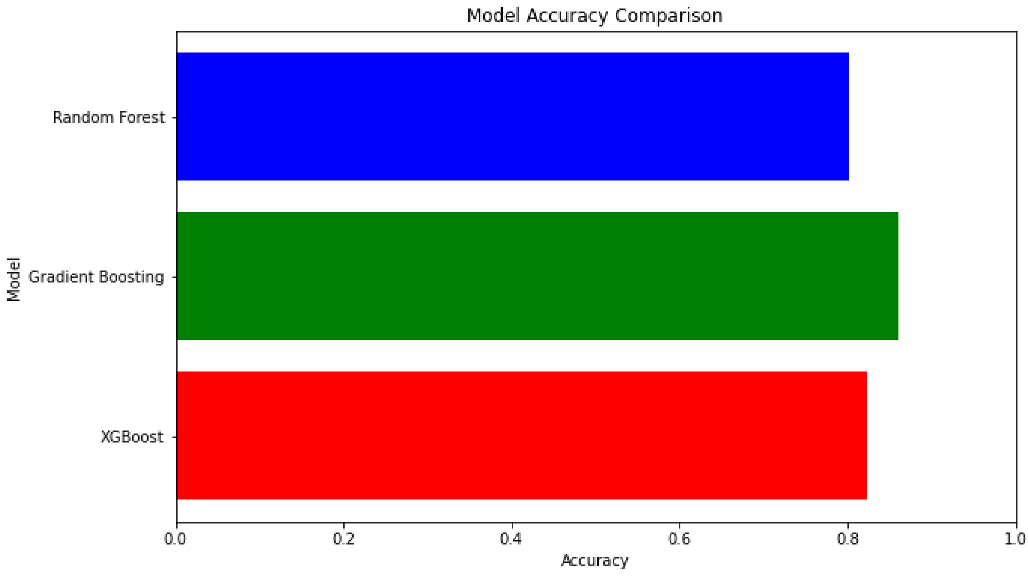
Together, these AI techniques support a more thorough and multifaceted approach to the identification and diagnosis of ASD, providing the opportunity for earlier, more precise, and individualized therapies.

Model Comparison:

Based on accuracy and performance criteria, the Random Forest, Gradient Boosting, and XGBoost models are compared to show their different advantages and disadvantages. With an accuracy of 80.15%, the Random Forest model showed strong recall for negative cases (90%) but lower precision (68%), suggesting some false positives. Its recall is only 74%, but its precision for positive cases is good at 93%. This model works well with big datasets and offers insightful information about the significance of features; nevertheless, it may overfit with too many trees and has marginally poorer overall accuracy.

Among the models, the Gradient Boosting model has the highest accuracy (86.03%). For positive examples, it performs admirably (accuracy 93%, recall 83%) and balances itself well (precision 77%, recall 90%) for negative instances. This model substantially lowers false positives as well as false negatives and is quite accurate. It does, however, require precise hyperparameter adjustment and is computationally demanding, which can be difficult.

XGBoost provides good performance, 82.35% accuracy. For positive examples, it retains great precision (88%) and memory (82%), whereas for negative occurrences, it performs well (74% precision, 83% recall). When parameters are adjusted properly, XGBoost which is well-known for its efficiency and scalability, generally achieves great performance. However, it can be resource-intensive and difficult to tune.



In conclusion, each model has advantages and disadvantages. However, the Gradient Boosting model stands out for having both balanced performance and excellent accuracy, which makes it a reliable option for real-world applications.

Implementation in Real- World Applications:

With the best accuracy and most balanced performance metrics of the three models, the Gradient Boosting model is especially well-suited for practical applications where dependability is essential. Applied to social issues, like helping people with autism, this kind of paradigm could transform early diagnosis and tailored intervention techniques. For instance, the approach may assist medical professionals in identifying early indicators of autism, allowing for prompt and customized interventions, by evaluating behavioral data, communication patterns, and other pertinent variables.

These models can help educators create individualized lesson plans that meet the special needs of kids with autism, improving their educational experience and results. Additionally, the model can be included into diagnostic tools in the healthcare industry to offer second opinions or screen huge populations for early indications of autism, which would help with resource allocation and focused support.

Researcher Prespective:

From the standpoint of a researcher, these models offer strong instruments for addressing a range of societal issues, including the requirements of people with autism. These models—Gradient Boosting in particular—have excellent accuracy and precision, which makes them useful for applications that need accurate classification and prediction. For example, these models can be used to create tailored education plans, assistive technologies, and early diagnosis tools for autism. These models can assist medical practitioners in early diagnosis and intervention, which are vital to enhancing the quality of life for people with autism. They do this by precisely recognizing patterns and generating predictions.

In summary, all three models have useful features, but the Gradient Boosting model is the most promising for practical use due to its balanced performance and higher accuracy, especially when it comes to helping people with autism and tackling social issues. We can improve many people's quality of life by using these cutting-edge machine learning approaches to increase diagnostic accuracy and personalize interventions.

Conclusion:

In the field of medical diagnostics, developing and use of machine learning models for the early diagnosis of Autism Spectrum Disorder (ASD) represents a major breakthrough. For the purpose of detecting ASD, this study analyzed face traits using three machine learning models: Random Forest, Gradient Boosting, and XGBoost. The Gradient Boosting model proved to be the most accurate, obtaining an astounding 86.03% accuracy rate, while the other models shown differing degrees of effectiveness and precision. With accuracies of 82.35% and 80.15%, respectively, the XGBoost and Random Forest models likewise demonstrated strong performance.

With its balanced precision and recall, the Gradient Boosting model performs better than other models, indicating its promise as a dependable method for early ASD detection. Through the use of advanced machine learning and picture classification methods, these models are able to recognize early indications of ASD, possibly even prior to the development of traditional symptoms. This capacity is essential for enabling prompt interventions and guaranteeing that kids who could have ASD get the care and assistance they require as soon as feasible.

The project aims to give every person thorough developmental examination in addition to improving the effectiveness of the diagnostic procedure. This method's automated, non-invasive design makes it possible to incorporate it easily into educational and medical environments, increasing its impact. Better developmental outcomes for children with ASD can result from giving doctors a trustworthy and impartial tool that speeds up the diagnostic process and supports decision-making.

The results of this study highlight how combining several datasets and using cutting-edge machine learning methods can enhance the performance and usefulness of the model. This method can greatly shorten the time needed to diagnose ASD and encourage early intervention, both of which are essential for enhancing the long-term results for those with ASD.

To sum up, the development of these machine learning models represents a significant advancement in the early diagnosis of autism. This research has the potential to change the support system for people with ASD by providing a dependable, practical, and easily available tool for use in both medical and educational situations. The utilization of these models has the potential to improve developmental pathways, facilitate timely interventions, and eventually raise the standard of living for individuals affected by the condition.

REFERENCES:

- https://www.theguardian.com/society/2024/mar/04/uk-increase-autism-diagnoses-neurodiversity#:~:text=Increases%20in%20diagnoses%20have%20been,autism%20spectrum%20disorder%20(ASD)
- https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/9711826
- https://thesai.org/Downloads/Volume15No2/Paper_64-Automated_Detection_of_Autism_Spectrum_Disorder.pdf
- https://ezproxy.herts.ac.uk/login?url=https://www.mdpi.com/2218-273X/14/1/48/pdf?version=1704182852
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8232415/
- https://www.irjcs.com/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9577321/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10264444/
- https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/10178739
- https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/10220851
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9958793/
- https://doaj.org/article/c4c7a1e9f16f4a8abe8556ed1c0679cf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8440949/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9533131/

- https://ezproxy.herts.ac.uk/login?url=https://www.mdpi.com/2078-2489/15/3/133/pdf?version=1709115402
- https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/9761040
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7824115/
- https://discovery.ucl.ac.uk/id/eprint/10166285/2/Topriceanu_Machine%20Learning%20with%20Neuroimaging%20Data%20to%20Identify%20Autism%20Spectrum%20Disorder%20%20A%20Systematic%20Review%20and%20Meta%20Analysis.pdf
- https://doaj.org/article/2d59ba50369c4ece8400649238c0b94e
- http://arxiv.org.ezproxy.herts.ac.uk/pdf/2206.11233
- https://link-springer-com.ezproxy.herts.ac.uk/article/10.1007/s40489-024-00435-4
- https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/10170707
- https://doaj.org/article/0ce674c9084b41e09827020e71d0f8bd
- https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/10481550
- https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/9989304

Appendix:

Below code is developed for project

Student ID: 21066861

Title: Vision for Understanding: Analysis of Facial Features in Autism Detection

Code:

```
import pandas as pd


hog_eye_region_path = r'C:\Users\SANSHIYA\Downloads\Autism Face Biometrics (AutFBio)\Autism Face Biometrics (AutFBio)\AutFBio\HOG-eye-region.csv'
hog_whole_face_path = r'C:\Users\SANSHIYA\Downloads\Autism Face Biometrics (AutFBio)\Autism Face Biometrics (AutFBio)\AutFBio\HOG-whole-face.csv'


hog_eye_region = pd.read_csv(hog_eye_region_path)
hog_whole_face = pd.read_csv(hog_whole_face_path)


#rename the coloumn
hog_eye_region = hog_eye_region.rename(columns={hog_eye_region.columns[0]: 'id'})
hog_whole_face = hog_whole_face.rename(columns={hog_whole_face.columns[0]: 'id'})
```

```
print(hog_eye_region.head())
print(hog_whole_face.head())

# Finding and merging common id from both dataset
Find_ids = set(hog_eye_region['id']) & set(hog_whole_face['id'])
print(f"Number of common IDs: {len(Find_ids)}")
hog_eye_region = hog_eye_region[hog_eye_region['id'].isin(Find_ids)]
hog_whole_face = hog_whole_face[hog_whole_face['id'].isin(Find_ids)]
merged_data = pd.merge(hog_eye_region, hog_whole_face, on='id', how='inner')

print("Merged Data Shape:", merged_data.shape)
print(merged_data.head())

# checking the target variable 'autism'
if 'autism_x' in merged_data.columns:
    merged_data = merged_data.rename(columns={'autism_x': 'autism'})
    merged_data = merged_data.drop(columns=['autism_y'])
elif 'autism_y' in merged_data.columns:
    merged_data = merged_data.rename(columns={'autism_y': 'autism'})
    merged_data = merged_data.drop(columns=['autism_x'])
else:
    raise ValueError("'autism' column not found in the merged dataset")

# Fill all missing values using median function
merged_data = merged_data.fillna(merged_data.median())

if 'id' in merged_data.columns:
    merged_data = merged_data.drop(columns=['id'])
X = merged_data.drop(columns=['autism'])
```

```
y = merged_data['autism'].values.ravel()
print("Shape of X:", X.shape)
print("Shape of y:", y.shape)
import matplotlib.pyplot as plt
import seaborn as sns

#visualise the autism rate in dataset
plt.figure(figsize=(8, 6))
ax = sns.countplot(x=y)
for p in ax.patches:
    ax.annotate(f'{int(p.get_height())}', (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='center', xytext=(0, 10), textcoords='offset points')
plt.title('Distribution of Autism in Dataset')
plt.xlabel('Autism')
plt.ylabel('Count')
plt.show()
# checking correlation using Heatmap
plt.figure(figsize=(12, 10))
correlation_matrix = merged_data.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()

import matplotlib.pyplot as plt
# Get feature importances
feature_importances = model.feature_importances_
feature_names = X.columns
feature_importances_df = pd.DataFrame({'Feature': feature_names, 'Importance': feature_importances})
feature_importances_df = feature_importances_df.sort_values(by='Importance', ascending=False)


# Print the top 10 most important features
print("Top 10 most important features:")
print(feature_importances_df.head(10))
plt.figure(figsize=(10, 6))
plt.title("Top 10 Feature Importances")
plt.barh(feature_importances_df['Feature'].head(10), feature_importances_df['Importance'].head(10))
plt.gca().invert_yaxis()
plt.xlabel("Importance")
plt.ylabel("Feature")
plt.show()
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import matplotlib.pyplot as plt
import seaborn as sns


# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


# Initialize and train the Random Forest model
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
```

```
# Print evaluation metrics
print(f"Accuracy: {accuracy}")
print("Confusion Matrix:")
print(conf_matrix)
print("Classification Report:")
print(class_report)

# Plot the confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", xticklabels=['Non-Autism', 'Autism'], yticklabels=['Non-Autism', 'Autism'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()from sklearn.ensemble import GradientBoostingClassifier

# Initialize and train the Gradient Boosting model
gb = GradientBoostingClassifier(random_state=42)
gb.fit(X_train, y_train)
y_pred_gb = gb.predict(X_test)
accuracy_gb = accuracy_score(y_test, y_pred_gb)
conf_matrix_gb = confusion_matrix(y_test, y_pred_gb)
class_report_gb = classification_report(y_test, y_pred_gb)

# Print evaluation metrics
print(f"Gradient Boosting Accuracy: {accuracy_gb}")
print("Gradient Boosting Confusion Matrix:")
print(conf_matrix_gb)
print("Gradient Boosting Classification Report:")
print(class_report_gb)

# Plot the confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix_gb, annot=True, fmt="d", cmap="Blues", xticklabels=['Non-Autism', 'Autism'], yticklabels=['Non-Autism', 'Autism'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Gradient Boosting Confusion Matrix')
plt.show()
from xgboost import XGBClassifier

# Initialize and train the XGBoost model
xgb = XGBClassifier(random_state=42)
xgb.fit(X_train, y_train)
y_pred_xgb = xgb.predict(X_test)
accuracy_xgb = accuracy_score(y_test, y_pred_xgb)
conf_matrix_xgb = confusion_matrix(y_test, y_pred_xgb)
class_report_xgb = classification_report(y_test, y_pred_xgb)

# Print evaluation metrics
print(f"XGBoost Accuracy: {accuracy_xgb}")
print("XGBoost Confusion Matrix:")
print(conf_matrix_xgb)
print("XGBoost Classification Report:")
print(class_report_xgb)

# Plot the confusion matrix
```

```
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix_xgb, annot=True, fmt="d", cmap="Blues", xticklabels=['Non-Autism', 'Autism'], yticklabels=['Non-Autism', 'Autism'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('XGBoost Confusion Matrix')
plt.show()

# Create the horizontal bar chart
import matplotlib.pyplot as plt
model_names = ['Random Forest', 'Gradient Boosting', 'XGBoost']
accuracies = [accuracy, accuracy_gb, accuracy_xgb]
plt.figure(figsize=(10, 6))
plt.barh(model_names, accuracies, color=['blue', 'green', 'red'])
plt.xlim(0, 1)  # Ensure the x-axis ranges from 0 to 1
plt.xlabel('Accuracy')
plt.ylabel('Model')
plt.title('Model Accuracy Comparison')
plt.gca().invert_yaxis()  # Invert y-axis to have the highest accuracy at the top
plt.show()
```