2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application

Efficient Comorbidity Analysis in Brain Disorders Reveals Better Diagnosis

Menglu Li*, Maryam Amirizaniani** , Rasha Kashef*
*Electrical, Computer, and Biomedical Engineering
**Ted Rogers School of Management
Ryerson University
Toronto, Canada
{menglu.li, maryam.amirizaniani, rkashef}@ryerson.ca

***Abstract*—Nowadays, brain disorders are gaining momentum faster than ever. Early detection of these disorders would be helpful in the treatment process. Also, detecting some comorbid brain disorders would be expensive and time-consuming. With advancements in machine learning (ML) and Artificial intelligence, these brain disorders and their comorbidities can be detected in the early stage. Different techniques of machine learning are used to detect Autism Spectrum Disorder (ASD), Attention deficit hyperactivity disorder (ADHD), Intellectual Disability (ID), and other brain disorders. This paper focuses on predicting ASD, ADHD, ID, and their Comorbidities via multi-stage analytical and prediction modelling. The first stage involves efficient data pre-processing. The next stage is a comorbidity analysis phase via logistic regression. In this analysis, logistic regression was applied to recognize health-related variables which are associated with ASD+ADHD+ID. These variables are Vision Test, Brain Injury, Anxiety, Down Syndrome, Blood Disorder, and Cystic Fibrosis. In the third stage, machine learning methods predict ASD, ADHD, ID for better diagnosis. For this purpose, SVM, KNN, and MLP are used. To evaluate these models, accuracy, precision, recall, and F1-score are selected.**

Keywords— Comorbidities, Brain Disorders, Comorbidities Analysis, Machine learning, Diagnosis, Evaluation.

I. INTRODUCTION

Comorbidity usually is associated with two or more illnesses or disorders, which may lead to worse health conditions, more complex clinical diagnoses, and treatment. Comorbidity in patients with brain disorders is prevalent [3]. Autism Spectrum Disorder (ASD) affects patients' social communication, imagination, and behaviour [1] with rising prevalence [2]. ADHD is a common neurodevelopmental disorder in childhood and lasts until adulthood. Children with ADHD have trouble paying attention, controlling impulsive behaviours, or being overly active. Comorbidity analysis in individuals with ASD or ADHD has received significant attention in the past few years to identify the relevant set of disorders associated with both disorders [3] for better treatment and resource allocations. In [4], a conducted research used health records of around 14000 patients to study the comorbidities of ASD with other diseases. To do so, chi-square statistics was employed as the method of this study. The result of this study shows that ASD has comorbidity with epilepsy, schizophrenia, inflammatory bowel disease, bowel disorders, CNS/cranial anomalies, diabetes mellitus type I, muscular dystrophy, and sleep disorders [4]. These are not the only diseases and disorders which are associated with ASD. Stevens et al. [5] showed that in children with ASD, the rate of comorbidity of ASD + ADHD is 42%, and the rate of comorbidity of ASD + ADHD +ID in children with ASD is 17%. Children with ASD/ADHD alone or with the comorbidity of ASD and ADHD have symptoms of diseases at different levels. Moreover, Goldin et al. [6] proposed that while Children with ADHD disorder show Tantrum behaviours, children with ASD disorders have higher Tantrum behaviours than children with ADHD. Also, tantrum behaviours were observed at relatively high levels in children with comorbid ASD and ADHD. These tantrum behaviours are included as they quickly become upset or angry, facing difficulties performing their tasks. From the research result of Jang et al. [7], children with ASD or ADHD alone have lower rates of psychopathology symptoms than individuals who have comorbid ASD and ADHD. In other words, children with comorbid ASD + ADHD have more problematic psychiatric symptoms like conduct behaviour problems, worry/depressed symptoms, avoidant behaviour, and tantrum behaviours. Moreover, finding the right treatments for individuals who have comorbid ASD+ADHD combined with another psychological problem can be more complicated. Fernell et al. [8] addressed the importance of ASD and ADHD early diagnosis due to the overlapping symptoms of their comorbidities, which would result in misdiagnosis and, subsequently, a failed treatment plan.

To the best of our knowledge, no studies have been performed on the National Survey Children Health 2018 (NSCH) [9] to provide a comprehensive comorbidity analysis of the three brain disorders, ASD, ADHD, and ID, for efficient diagnosis using various machine learning prediction models. In this paper, we propose a multi-stage diagnosis procedure that achieved better accuracy than the state-of-the-art methods to reduce the risk of misdiagnosis. In the first stage, we recognize comorbidities in ASD, ADHD, ID or combined by providing an efficient association analysis among all the common correlated comorbidities in the three brain disorders. Our comorbidity analysis reveals the underlying relationship among the comorbid symptoms. This stage can identify the key variables that are related to ASD, ADHD, ID, and their comorbidities. These selected variables will be passed to the next stage of our model and be used to provide early diagnosis of each disorder with better prediction accuracy. For this purpose, we use various machine learning methods to classify each disorder efficiently. The main contribution of the proposed model is to analyze and identify the major symptoms of the brain disorders and their comorbidities, which broaden

978-1-6654-9457-1/21/$31.00 ©2021 IEEE

DOI 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00189

the understanding of the disorder prediction process. The proposed model is assessed using multiple quality measures as precision, recall, F-score, and Accuracy. Experimental results on the NSCH dataset reveal that the accuracy of predicting ASD is up to 97.19%, while ADHD was predicted with an accuracy of up to 89.78%. The ID disorder has the highest accuracy of up to 99.07%. We can also observe that the dual disorders ASD+ADHD, ASD+ID, and ADHD+ID are diagnosed with an accuracy of 99.72%, 91.52%, and 98.8%, respectively. The prediction of the triple disorders has a prediction accuracy of 88.2%, which is very close to the accuracy of the multi-class problem (88.18%).

The rest of this paper is organized as follows: Section 2 discusses the literature review. In section 3, proposed methodologies are discussed, and experimental Analysis and Results are introduced and presented in Section 4. Finally, the conclusion and future directions are given in Section 5.

II. LITERATURE REVIEW

This section outlines related work on machine learning methods in healthcare, comorbidity Analysis, and Brain disorders.

A. *Machine Learning in healthcare*

Because of the power of Machine learning (ML) to process vast amounts of data and extract features, ML-related techniques have been applied to the healthcare industry for applications, such as diagnosis and medical events prediction. Nowadays, machine learning models are used abroad in medicine and biology for cancer, discoveries of a new novel, genomics, and imaging data interpretation [10]. Raita et al. [11] applied multiple ML techniques, Lasso regression, random forest, gradient boosted decision tree, and deep neural network to prove that machine learning algorithms can enhance clinicians' triage decision making and prioritize critical patients. The ML-based techniques outperform the traditional Emergency Severity Index (ESI) algorithm. Ramkumar et al. [12] applied an artificial neural network-based algorithm to predict inpatient information. The proposed model uses 15 preoperative features as inputs, including age, gender, type of admission, and the number of associated diagnoses. The predicted result of inpatient information achieves the AUC score of around 80%, demonstrating the capability of ANN-based algorithms with validity, responsiveness, and reliability in predicting inpatient information metrics and patient-specific case complexity. Rajkomar et al. [13] suggested three other Deep Learning (DL) methods based on LSTM, TANN, and neural networks with boosted time-based decision stump to predict patients' medical events, such as in-hospital mortality, readmission rate, and length of stay, which also demonstrates the ability of ML algorithms to extract information from Electronic Health Record (EHR). Yang et al. [14] proposed a hybrid ML-based method for classifying antigens as cancer or not cancer. The testing antigens data passes through the Self-organizing Map (SOM) classifier and the Recursive maximum contrast trees (RMCT) classifier. If the results are not the same, the data will be passed to the third classifier, the Parallel Self-organizing Hierarchical Neural Network (PSHNN).

B. *Machine Learning in Comorbidity Analysis*

Boytcheva et al. [15] proposed a ML-based approach to extract potential comorbidity patterns in the big collection of outpatient records. Firstly, text mining tools are applied to convert the free text in EHR to structured data. The MIxCO algorithm performed comorbidity mining by finding the maximal frequent patterns from the structured text data. Zhang et al. [16] also predicted the comorbid risk by proposing heterogeneous Convolutional Neural Networks (HCNN). The proposed HCNN algorithm extracted the temporal relationship between the heterogeneous diagnoses from patients' EHR data into a graph. The temporal intervals computed the edges in the graph. Then, a five-layer CNN is applied to the graph data. Zhang et al. [16] showed that a graph structure could represent the relationship between comorbidities. Farran et al. [17] worked on the diagnosis of comorbidity. They proposed a two-stage supervised classification algorithm to predict the diagnosis for diabetes, hypertension, and comorbidity. The first stage applies the Support Vector Machines to classify diabetes in the general population, and the second stage uses k-NN techniques to classify diabetes in the hypertensive population. The experiment result indicated that the proposed two-stage algorithm has a 10% higher prediction accuracy rate than only one classification algorithm. Dashtban et al. [18] recognized early readmission for comorbidity patients by applying the Generative Adversarial Network (GAN), which has a solid ability to handle noisy data and missing values. Machine learning models can help to make improvements in the comorbidity of medicine. Wang et al. [19] proposed a DL-based model, PPC, which combines patient information, and medical ontologies to recommend a personalized prescription for comorbidity. This model can learn the patient characteristics with the MLP model, and the result shows that the Micro-AUC of the PPC model is 93.1%. To test different ML-based techniques' ability to analyze comorbidity, Zolbanin et al. [20] applied multiple ML-based algorithms to predict overall survivability in the comorbidity of cancers. The pre-processed dataset of cancer patients is passed separately into Artificial Neural Networks (ANN), Logistic Regression (LR), Random Forest (RF), and Decision tree. The result shows that deep learning techniques generally perform better than traditional classification techniques.

C. *Machine Learning in Brain Disorders*

A few ML-related studies focused on analyzing ASD / ADHD/ID or their comorbidities. Doshi-Velez et al. [21] applied unsupervised clustering on the EHR to investigate patterns of co-occurrence of medical comorbidities in ASDs. Some of the commodities for ASD include seizures, psychiatric illness, and complex multisystem disorders, including auditory and gastrointestinal disorders. These Comorbidities do not occur equally in patients. Usta et al. [22] used predictive factors of ASD on Naive Bayes, Generalized Linear Model, Logistic Regression, Decision Tree to diagnose ASD patients. These factors include parental age, birth weight, pre-treatment IQ, sociodemographic variables, comorbid psychiatric disorders, language skills, etc. The research of Bishop-Fitzpatrick et al. [23] tried to identify the comorbidity caused by ASD. They applied random forest on

ICD-9 codes, V-codes, and E-codes in EHRs and received a high level of detection accuracy. Asif et al., in 2020, used a machine learning model to predict ASD based on patients' clinical profiles, disrupted biological processes, and brain genes. A Naive Bayes classifier was used to predict ASD patients. This model has a precision of 0.82 and a low recall of 0.39 [24]. Alkoot et al. [25] used a 4-level machine learning model to detect ASD according to patients' genetic information. This 4-level machine learning model is included as a k-nearest neighbour, 1-nearest neighbour, back-propagation neural network, and support vector machine classifiers. Gori et al. [26] studied an SVM model to analyze brain morphometry of children with ASD and classified the brain features and brain regions associated with ASD. Küpper et al. [27] used an SVM model to predict different categories of behavioural features associated with ASD, such as stereotyped/idiosyncratic use of words or phrases, conversation, emphatic or emotional gestures, unusual eye contact, facial expressions.

Stevens et al. [28] identified the behavioural phenotypes of ASD using the unsupervised machine learning method. The sample data of children are divided into subgroups by a Gaussian Mixture model. Then, the behavioural phenotypes are examined through Hierarchical Agglomerative Clustering. Cantin-Garside et al. [29] used KNN and SVM to classify ASD patients. The highest achieved accuracy in KNN and SVM models is 99.1%. The mean accuracy in these models is 93% [29]. Maenner et al. [30] used the random forest model to predict the ASD status of 8-years old children from the 2008 Georgia ADDM site. Detecting ASD in clinics is expensive and time-consuming. Omar et al. [31] developed a mobile application to detect ASD based on machine learning models. In this mobile application, they merged Random Forest Cart (RF-CART) and Random Forest ID3 (RF-ID3) to detect ASD patients [31].

The study of Mueller et al. [32] classified ADHD adults using a support vector machine (SVM). This is the first attempt to use non-linear machine learning methods in the context of clinical groups. The independent event-related potentials (ERP) are used as features of SVM. Oztoprak et al. [33] proposed a model called SVM-RFE (Support Vector Machine-Recursive Feature Elimination). The SVM-RFE builds upon the research result of Mueller et al. [34], which also accepted the ERP as a feature.

Zhang et al. [33] proposed a longitudinal recurrent neural network (RNN) model with the Long Short-Term Memory (LSTM) to predict the comorbidity of ADHD, especially for the Substance Use Disorders (SUDs) using Swedish registry data. The RNN model can predict the SUDs ten years before the earliest diagnosis based on the data of ADHD patients. Peng et al. [34] used SVM and ELM machine learning models to find an effective and accurate diagnosis for ADHD. They achieved an accuracy of 90.18% for ELM and 84.73% for SVM based on the MRI of patients, cortical features, and brain segments [35]. Kim et al. [36] proposed an SVM model to predict ADHD based on pre-treatment demographic, clinical questionnaire, neuroimaging, environmental, neuropsychological, and genetic information.

Duda et al., 2016 [37] also worked on ASD diagnosis, which used six machine learning models to distinguish symptoms of ASD from ADHD based on the behaviour of patients. The main goal of this study is to use ML models to speed up the distinction of ASD and ADHD processes. In another research, Duda et al., 2017 [38] used a crowdsource dataset. They applied SVC, Logistic regression with Lasso regularization, Logistic regression with Ridge regularization, LDA, and Elastic Net (ENet), to identify to distinguish symptoms of ASD from ADHD. Aggarwal & Singh [39] researched diagnosing ID based on speech features by machine learning models. This research used 4 machine learning models as KNN, SVM, RBFNN, and LDA. The highest accuracy achieved among these models was 96% in the RBFNN model [39]. Bertoncelli et al. [40] used a predictive machine learning method to identify the ID's factors in teenagers with cerebral palsy. Based on a logistic regression model, poor manual abilities, gross motor function, and type of epilepsy are significantly associated with intellectual disability. Few research studies have focused on using Machine Learning and comorbidity analysis in ASD, ADHD, and ID.

Cordova et al. [41] studied executive function, shared or distinct across ADHD and ASD, using a supervised random forest and functional random forest to observe an executive function like hyperactivity and inattention. It is worth mentioning that some of the genes related to ASD are mutual with ID genes, and according to these genes, ASD + ID patinas can be detected. Kou [42] aimed to diagnose ASD + ID based on genes and additional functional information such as protein. Their method was based on an SVM model; they predicted ASD, ID, and ASD +ID with accuracy above 80% [42].

III. The Proposed Methodologies

The proposed methodologies involve four phases: (1) pre-processing, (2) comorbidity analysis, (3) diagnosis, and (4) assessment, as shown in Fig.1.

- All missing cells in the dataset are imputed in the pre-processing phase, with the average value for each missing feature in the dataset. We initially identified ASD, ADHD, ID as dependent variables. Based on these columns, we define four dependant variables as ASD+ADHD, ASD+ID, ADHD+ID, and ASD+ADHD+ID variables. We selected all health-related variables in this dataset as the significant independent variables to perform the next stage of analysis, which is comorbidity analysis.
- In the comorbidity analysis, we use logistic regression models to explore ASD, ADHD, and ID comorbidities. Firstly, we used logistic regression to analyze diseases and disorders which are statistically significant for ASD, ADHD, and ID. Then, we apply another set of logistic regression models to explore diseases and disorders associated with ASD, ADHD, ID, and ASD+ADHD+ID. After completing the comorbidity analysis, significant features related to disorders are selected. These selected variables can represent the characteristics of the disorders.

- In the diagnosis stage, we use three machine learning methods to predict each disorder and its comorbidity using the selected significant features. The machine learning models are Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Multi-Layer Perceptron (MLP). SVM is a supervised machine learning algorithm that tries to find a hyperplane in an N-dimensional space to classify the data instances. This algorithm performs well on binary classification problems. The resulting hyperplane has the maximum distance between the data instances of both classes. Also, the SVM supports the Radial basis function as its kernel function so that the model can solve the non-linearly separable problem. KNN is a simple and widely used machine learning technique for classification. The model assigns a label to a new data instance by finding the k most similar instances in the training set. The most common label among these k similar instances will be the predicted label to assign. KNN performs well when the training set is large enough. However, the prediction result may be affected by the outliers in the dataset. MLP is a deep learning technique, which is a class of Feedforward Neural networks (FFNN). This technique is widely used in non-linear classification problems. The structure of MLP contains an input layer, a hidden layer, and an output layer. The hidden layer can approximate any continuous function. Therefore, MLP performs well when there is a complex relationship between features and its label.
- In the assessment phase, we use four evaluation metrics [43],[44],[45] to measure each algorithm's classification performance: accuracy, precision, recall, and F1-score. Accuracy shows the number of correct predictions in the process of classification. However, this measure is not useful when the dataset is imbalanced. Therefore, we include precision, recall, and F1-score to observe the number of true positive and true negative predictions. Multiple evaluation metrics can provide more accurate and complete information about classification performance.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

A. Datasets

The dataset used in this paper is obtained from the National Survey Children Health 2018 (NSCH), which has information about the health and well-being of children in ages 0 to17 years old [42]. The NSCH gathered information on just one child from each household. The NSCH provided health information for 30,530 children in 2018. After filtering variables and removing blank cells for this study, this dataset has around 30,000 records. In this paper, we address three dependant variables, including ASD, ADHD, and ID. There are 24 health-related variables from the NSCH datasets selected for this study as the independent variables. They are Dental Service, Vison Test, Breathing Difficulty in the past 12 months, Swallowing Difficulty in past 12 months, Stomach Difficulty in past 12 months, Deafness, Blindness, Allergies, Asthma, Brain Injury, Cerebral Palsy, Diabetes, Epilepsy, Heart Condition, Headaches, Tourette Syndrome, Anxiety, Depression, Down Syndrome, Blood Disorder, Blood Disorder, Cystic Fibrosis. These selected independent variables will help identify the comorbid disease and disorders of ASD, ADHD, and ID.

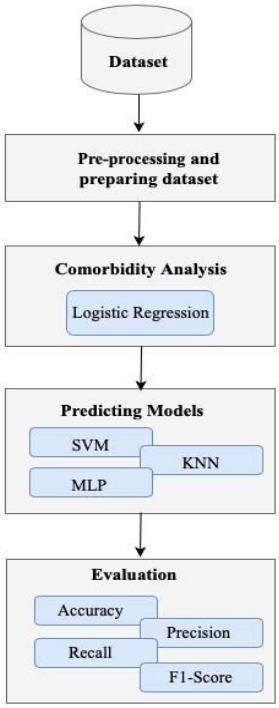

Fig.1. Comorbidity Analysis and Diagnosis.

B. Comorbidity Analysis

This stage aims to explore the dependant variables that can be removed and the predictor variables that significantly identify the comorbidities' characteristics. There is no multicollinearity problem in this study. The significant variables will be analyzed to find comorbidities with ASD, ADHD, and ID. The selection of potential cofounders of ASD, ADHD, and ID and empirical evidence is considered. For this empirical evidence, a regression was applied to identify cofounders with P-value equal to or less than 0.05. So, logistic regression, LRModel 1, was fitted to measure significant variables associated with each of the three dependent variables (ASD, ADHD, and ID). The independent variables with a p-value less than 0.05 are identified, marked as * in Table 1. For comorbidity analysis purposes, these statically significant variables were placed in another regression against independent variables (ASD, ADHD, and ID). Logistic regression analyses were done on dependent and independent variables to analyze the relationship between specific comorbidities and the three-dependent variables (ASD, ADHD, and ID). After considering their interactions with other variables, this regression aims to identify significant variables associated with ASD, ADHD, and ID.

A Logistic regression model, LRModel 2, was performed to specify the relationship between ASD as a dependent variable and significant variables related to ASD. TABLE II shows the result of this regression. The variables with a p-value less than 0.05 are identified as substantial variables for

ASD. As shown in TABLE II, Vision Test, Breathing Difficulty, Swallowing Difficulty, Stomach Difficulty, Deafness, Blindness, Brain Injury, Headaches, Tourette Syndrome, Anxiety, Down Syndrome, Blood Disorder, and Cystic Fibrosis significantly associated with ASD. The correlation between ASD and significant variables is calculated such that we concluded that there is no strong correlation between these variables. This highlights the association of significant variables with ASD.

TABLE I. Logistic Regression and Significance testing

	ASD		***ADHD***		***ID***	
	COEF	***P-Value***	***COEF***	***P-Value***	***COEF***	***P-Value***
Dental Service	0.004	0.000*	0.008	0.000*	0.004	0.000*
Vision Test	0.007	0.000*	0.012	0.000*	0.004	0.000*
Breathing Difficulty	-0.021	0.006*	0.006	0.638	-0.022	0.002*
Swallowing Difficulty	-0.014	0.016*	0.002	0.824	0.003	0.556
Stomach Difficulty	0.02	0.000*	0.008	0.311	0.013	0.006*
Deafness	-0.017	0.003*	0.004	0.68	0.021	0.000*
Blindness	0.049	0.000*	0.006	0.584	-0.02	0.001*
Allergies	-0.008	0.36	0.015	0.301	-0.04	0.551
Arthritis	-0.004	0.123	0.008	0.126	0.005	0.065
Asthma	0.0004	0.928	0.014	0.030*	0	0.803
Brain Injury	0.033	0.000*	0.061	0.000*	-0.01	0.013*
Cerebral Palsy	0.006	0.292	0.03	0.006*	0.014	0.017*
Diabetes	-0.012	0.054	0.025	0.018*	0.004	0.414
Epilepsy	0.003	0.596	0.028	0.563	0.007	0.27
Heart Condition	0.008	0.246	-0.012	0.363	0.009	0.208
Headaches	0.026	0.000*	-0.004	0.706	0.029	0.000*
Tourette Syndrome	0.059	0.000*	0.007	0.506	0.03	0.000*
Anxiety	0.03	0.000*	0.006	0.031*	0.016	0.007*
Depression	0.007	0.214	-0.002	0.829	0.041	0.000*
Down Syndrome	0.089	0.000*	0.051	0.000*	0.102	0.000*
Blood Disorder	0.148	0.000*	0.121	0.000*	0.251	0.000*
Cystic Fibrosis	0.08	0.000*	0.0419	0.002*	0.108	0.000*

To specify the relationship of ADHD as a dependent variable and significant variables related to ADHD, a logistic regression model, LRModel 3, was performed. TABLE III shows the result of this regression. According to this result, Dental Service, Vision Test, Asthma, Brain Injury, Cerebral Palsy, Diabetes, Epilepsy, Down Syndrome, Blood Disorder, and Cystic Fibrosis have a p-value less than 0.05, which means they are significantly associated with ADHD. The Pearson correlation between ADHD and significant variables presents that there is a weak correlation between these variables. This shows the association of significant variables with ADHD.

The last step of the comorbidity analysis is to specify the relationship of ID as a dependent variable and significant variables; a logistic regression, LRModel 4, was conducted. TABLE IV shows the result of LRModel 4. Based on TABLE IV, Dental Services, Vision Test, Breathing Difficulty, Stomach Difficulty, Blindness, Brain Injury, Cerebral Palsy, Headaches, Tourette Syndrome, Anxiety, Depression, Down Syndrome, Blood Disorder, and Cystic Fibrosis significantly associated with ID. The Pearson correlation between ID and significant variables shows that there is not a correlation between these variables.

TABLE II. Significance Testing: LRModel 2

	COEF	***P-VALUE***
Dental Service	0.006	0.15
Vision Test	0.009	0.000*
Breathing Difficulty	0.0103	0.000*
Swallowing Difficulty	0.0084	0.000*
Stomach Difficulty	0.0074	0.000*
Deafness	0.0103	0.000*
Blindness	0.0121	0.000*
Brain Injury	0.0192	0.000*
Headaches	0.0364	0.000*
Tourette Syndrome	0.0275	0.000*
Anxiety	0.0241	0.000*
Down Syndrome	0.057	0.000*
Blood Disorder	0.122	0.000*
Cystic Fibrosis	0.056	0.000*

TABLE III. Significance Testing: LRModel 3

	COEF	***P-VALUE***
Dental Service	0.0061	0.000*
Vision Test	0.0011	0.000*
Asthma	0.0069	0.000*
Brain Injury	0.0215	0.000*
Cerebral Palsy	0.0152	0.000*
Diabetes	0.193	0.000*
Anxiety	0.026	0.000*
Down Syndrome	0.061	0.000*
Blood Disorder	0.134	0.000*
Cystic Fibrosis	0.0611	0.000*

TABLE IV. Significance Testing: LRModel 4

	COEF	*P-VALUE*
Dental Service	0.006	0.000*
Vision Test	0.0098	0.000*
Breathing Difficulty	0.0098	0.000*
Stomach Difficulty	0.0072	0.000*
Deafness	0.009	0.147
Blindness	0.0117	0.000*
Brain Injury	0.0175	0.000*
Cerebral Palsy	0.0128	0.000*
Headaches	0.034	0.000*
Tourette Syndrome	0.0268	0.000*
Anxiety	0.0222	0.000*
Depression	0.0131	0.000*
Down Syndrome	0.0552	0.000*
Blood Disorder	0.1234	0.000*
Cystic Fibrosis	0.0555	0.000*

Based on results from Tables 1-4 and the correlation analysis, it can be concluded that the 6 features, ***Vision Test, Brain Injury, Anxiety, Down Syndrome, Blood Disorder, and Cystic Fibrosis***, are significant comorbidities with the three disorders ASD, ADHD, and ID. These features will then be used for the diagnosis and assessment phase discussed next.

C. Diagnosis Using Predication Modelling

After the Comorbidity analysis was performed, multiple predictive models based on machine learning were adopted to predict ASD, ADHD, or ID. We have used the SVM, KNN, and MLP classifiers to predict each disorder based on the six final variables, ***Vision Test, Brain Injury, Anxiety, Down Syndrome, Blood Disorder, and Cystic Fibrosis.***

We run two types of experiments: binary classification and multi-class classification. In the 2-class problem, we assume that if an individual has a disorder d, were

$$d \in \{ASD, ADHD, ID, ASD + ADHD, ASD + ID, ADHD + ID, ASD + ADHD + ID, None\}$$

Then the class label is one, and other classes are considered zero. We run 7 trials in this problem for each disorder. For the multi-class problem, we consider eight classes: $\{ASD, ADHD, ID, ASD + ADHD, ASD + ID, ADHD + ID, ASD + ADHD + ID,$ and $None\}$.

In both experiments, we used 70% as training and 30% as testing. The Accuracy, Precision, Recall, and F-score metrics are shown in Figures 2-5, respectively.

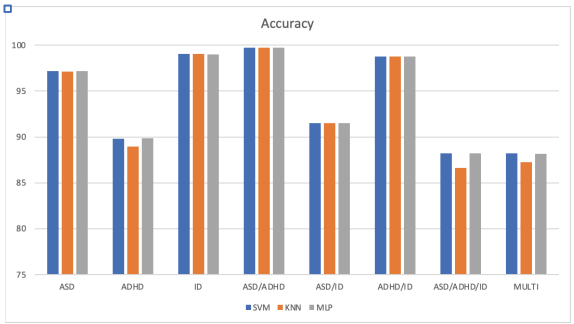

Fig. 2. SVM, KNN, MLP (Accuracy)

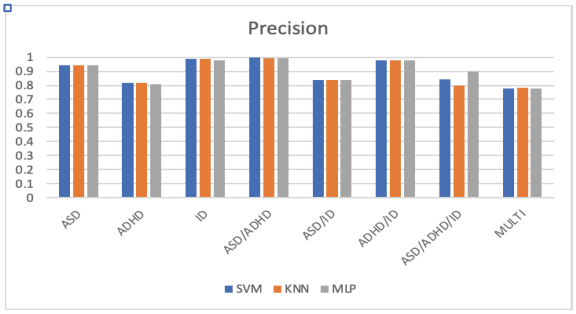

Fig. 3. SVM, KNN, MLP (Precision)

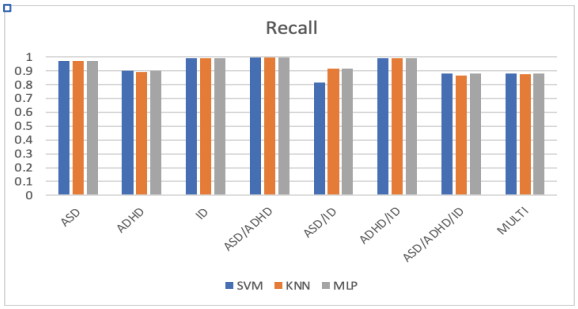

Fig. 4. SVM, KNN, MLP (Recall)

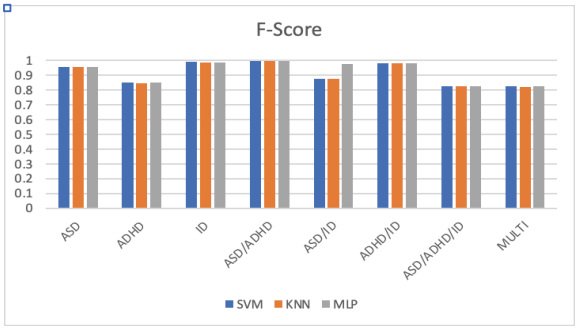

Fig. 5. SVM, KNN, MLP (F-score)

For the 2-class problem, it can be shown that the accuracy of predicting ASD is up to 97.19%, while ADHD was detected with an accuracy of up to 89.78%. The ID disorder has the highest accuracy of up to 99.07%. We can also observe that the dual disorders ASD+ADHD, ASD+ID, and ADHD+ID are diagnosed with an accuracy of 99.72%, 91.52%, and 98.8%, respectively. The prediction of the triple disorders has the lowest prediction accuracy of 88.2%, which is very close to the accuracy of the multi-class problem (88.18%). Figures 3 and 4 show that SVM, KNN, and MLP have similar performance in predicting ASD, ID, ASD+ADHD, ASD+ID, and ADHD+ID. KNN has the best performance in diagnosing ADHD, and MLP has the highest prediction for triple disorders. For the F-score, we can observe from Figure 6 that the KNN has the best performance for all disorders of up to 0.9966. It can be shown that the three prediction models achieve the same performance for the multi-class classification problem. It can be demonstrated that the comorbidity analysis has significantly enhanced the prediction accuracy of up to 99%.

As the KNN outperforms SVM and MLP in most cases, we have investigated the performance of the KNN using ASD marker only (7 variables: Breathing Difficulty, Swallowing

difficulty, stomach difficulty, Deafness, Blindness, Headaches, and Tourette syndrome), ADHD marker only (Dental Service, Asthma, Cerebral Palsy, and Diabetes), and ID markers only (Dental Service, Breathing Difficulty, stomach difficulty, Deafness, Cerebral Palsy, Blindness, Headaches, and Tourette syndrome).

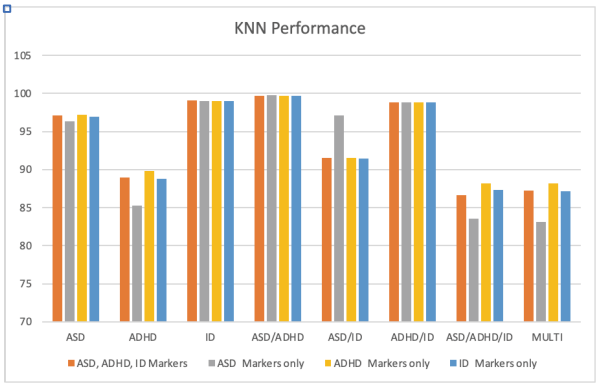

Fig. 6. KNN (Accuracy): Individual markers

In Fig.6, we can observe that the performance of the KNN for the 2-class problem had decayed for the ADHD when we used ASD markers or ID markers and increased with the ADHD markers. Similarly, the ASD markers have improved the accuracy of the KNN in predicting ASD/ID as there are five common markers in both disorders. Using the ASD marker only has caused a significant drop in the accuracy for predicting the triple disorders for the 2-class and the multi-class problems. While the ADHD markers only have enhanced the performance in both cases.

V. CONCLUSION AND FUTURE DIRECTIONS

Early detection of brain disorders is important in assessment and treatment processes. Symptoms of some disorders might overlap; thus, the diagnosis of these disorders would be challenging. This paper proposed a multi-stage methodology to predict brain disorders using comorbidity analysis. The explored disorders in this study are ASD, ADHD, ID, and their comorbidities. We concluded that Vision Test, Brain Injury, Anxiety, Down Syndrome, Blood Disorder, and Cystic Fibrosis are the common comorbidities in the three brain disorders. In this paper, we successfully deployed various machine learning models to diagnose ASD, ADHD, ID better or combined. These models are SVM, KNN, and MLP. Accuracy, precision, recall, and F-Score, are used to evaluate the adopted models. For Future research, we will expand our diagnosis models using deep learning with more layers or hybrid learning. Besides, we aim to use unsupervised machine learning to provide clusters of comorbidities and disorders for unlabelled datasets.

REFERENCE

[1] L. Waterhouse, “Rethinking autism: Variation and complexity.”, Academic Press, 2013.

[2] J. L. Matson and A. M. Kozlowski, "The increasing prevalence of autism spectrum disorders," Research in Autism Spectrum Disorders, vol. 5, (1), pp. 418-425, 2011.

[3] J. L. Matson and R. L. Goldin, "Comorbidity and autism: Trends, topics and future directions," Research in Autism Spectrum Disorders, vol. 7, (10), pp. 1228-1233, 2013.

[4] I. S. Kohane, A. McMurry, G. Weber, D. MacFadden, D., Rappaport, L., Kunkel, L., S. Churchill, S. “The comorbidity burden of children and young adults with autism spectrum disorders,” PloS One, 7(4), e33224-e33224, 2012.

[5] T. Stevens, L. Peng, and L. Barnard-Brak, “The comorbidity of ADHD in children diagnosed with autism spectrum disorder,” Research in Autism Spectrum Disorders, 31, pp.11-18. 2016.

[6] R. Goldin, J. Matson, K. Tureck, P. Cervantes, and J. Jang, “A comparison of tantrum behaviour profiles in children with ASD, ADHD and comorbid ASD and ADHD,” Research in Developmental Disabilities, 34(9), pp.2669-2675, 2013.

[7] J. Jang, J. Matson, L. Williams, K. Tureck, R. Goldinand P. Cervantes, P., “Rates of comorbid symptoms in children with ASD, ADHD, and comorbid ASD and ADHD,” Research in Developmental Disabilities, 34(8), pp.2369-2378, 2013.

[8] E. Fernell, M.A. Eriksson, and C. Gillberg,“Early diagnosis of autism and impact on prognosis: a narrative review,” Clinical epidemiology, 5, p.33, 2013.

[9] Child and Adolescent Health Measurement Initiative, Data Resource Center on Child and Adolescent Health website. The National Survey Children Health, 2018. Retrieved from https://www.childhealthdata.org/dataset/download?rq=9531

[10] E. Mossotto, JJ. Ashton, T. Coelho, RM. Beattie, D. MacArthur, S. Ennis, “Classification of pediatric inflammatory bowel disease using machine learning”. Sci Rep. 2017;7:2427. doi:10.1038/ s41598-017-02606-2.

[11] Y. Raita, T. Goto, M. Faridi, D. Brown, C. Camargo, and K. Hasegawa, “Emergency department triage prediction of clinical outcomes using machine learning models,” Critical Care, 23(1), 2019.

[12] P. Ramkumar, J. Karnuta, S. Navarro, H. Haeberle, R. Iorio, M. Mont, B. Patterson, and V. Krebs, “Preoperative Prediction of Value Metrics and a Patient-Specific Payment Model for Primary Total Hip Arthroplasty: Development and Validation of a Deep Learning Model,” The Journal of Arthroplasty, 34(10), pp.2228-2234.e1, 2019.

[13] A. Rajkomar, E. Oren, K. Chen, A. Dai, N. Hajaj, M. Hardt, P. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G., Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. Shah, A. Butte, M. Howell, C. Cui, G. Corrado, and J, Dean, “Scalable and accurate deep learning with electronic health records,” npj Digital Medicine, 1(1), 2018.

[14] J. Yang, M. Yang, Z. Luo, Y. Ma, J. Li, Y. Deng, and X. Huang, “A hybrid machine learning-based method for classifying the Cushing's Syndrome with comorbid adrenocortical lesions,” BMC Genomics, 9(Suppl 1), p.S23, 2008.

[15] S. Boytcheva, G. Angelova, Z. Angelov, and D. Tcharaktchiev, “Mining comorbidity patterns using retrospective analysis of big collection of outpatient records,” Health Information Science and Systems, 5(1), 2017.

[16] J. Zhang, J. Gong, and L. Barnes, “HCNN: Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records,” 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017.

[17] B. Farran, A. Channanath, K. Behbehani, and T. Thanaraj, “Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study,” BMJ Open, 3(5), p.e002457, 2013.

[18] M. Dashtban, and W. Li, “Predicting risk of hospital readmission for comorbidity patients through a novel deep learning framework,” In Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020.

[19] L. Wang, W. Zhang, X. He, and H. Zha, “Personalized Prescription for Comorbidity. In: Pei J., Manolopoulos Y., Sadiq S., Li J. (eds) Database Systems for Advanced Applications. DASFAA 2018. Lecture Notes in Computer Science, vol 10828,” Springer, Cham. 2018. https://doi.org/10.1007/978-3-319-91458-9_1

[20] H. Zolbanin, D. Delen, and A. Hassan Zadeh, “Predicting overall survivability in comorbidity of cancers: A data mining approach,” Decision Support Systems, 74, pp.150-161, 2015.

[21] F. Doshi-Velez, Y. Ge, and I. Kohane, “Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis,” Pediatrics, 133(1), pp.e54-e63, 2013.

[22] M. Usta, K. Karabekiroglu, B. Sahin, M., Aydin, A. Bozkurt, T. Karaosman, A. Aral, C. Cobanoglu, A. Kurt, N. Kesim, I Sahin, and E. Ürer, “Use of machine learning methods in prediction of short-term outcome in autism spectrum disorders,” Psychiatry and Clinical Psychopharmacology, 29(3), pp.320-325, 2018.
[23] L. Bishop-Fitzpatrick, A. Movaghar, J. S. Greenberg, D. Page, L. S. DaWalt, M. H. Brilliant, and M. R. Mailick, “Using machine learning to identify patterns of lifetime health problems in decedents with autism spectrum disorder,” Autism Research. 1120–1128, 2018.
[24] M. Asif, Martiniano, F. Hugo, A.R. Marques, J.X. Santos, J. Vilela, C. Rasga, G. Oliveira, F.M. Couto, and A.M. Vicente, "Identification of biological mechanisms underlying a multidimensional ASD phenotype using machine learning", Translational Psychiatry, vol. 10, no. 1, pp. 43-43, 2020.
[25] F. M. Alkoot and A. K. Alqallaf, "Investigating machine learning techniques for the detection of autism," International Journal of Data Mining and Bioinformatics, vol. 16, (2), pp. 141-169, 2016.
[26] I. Gori, A. Giuliano, F. Muratori, I. Saviozzi, P. Oliva, R. Tancredi, A. Cosenza, M. Tosetti, S. Calderoni, and A. Retico, "Gray Matter Alterations in Young Children with Autism Spectrum Disorders: Comparing Morphometry at the Voxel and Regional Level: Gray Matter Alterations in ASD Children", Journal of neuroimaging, vol. 25, no. 6, pp. 866-874, 2015.
[27] C. Küpper et al., "Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning," Scientific Reports, vol. 10, (1), pp. 4805-4805, 2020.
[28] E. Stevens, D. R. Dixon, M. N. Novack, D. Granpeesheh, T. Smith, and E. Linstead, “Identification and analysis of behavioural phenotypes in autism spectrum disorder via unsupervised machine learning,” International Journal of Medical Informatics, vol. 129, pp. 29–36, 2019.
[29] K.D. Cantin-Garside, Z. Kong, S.W. White, L. Antezana, S. Kim, and M.A. Nussbaum, "Detecting and Classifying Self-injurious Behavior in Autism Spectrum Disorder Using Machine Learning Techniques", Journal of autism and developmental disorders, vol. 50, no. 11, pp. 4039-4052, 2020.
[30] M.J. Maenner, M. Yeargin-Allsopp, K. Van Naarden Braun, D.L. Christensen, and L.A. Schieve, "Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder", PloS one, vol. 11, no. 12, pp. e0168224-e0168224, 2016.
[31] K.S. Omar, P. Mondal, N.S. Khan, M.R.K. Rizvi, and M.N. Islam, "A Machine Learning Approach to Predict Autism Spectrum Disorder", IEEE, pp. 1, 2019.
[32] A. Mueller, G. Candrian, J. D. Kropotov, V. A. Ponomarev, and G.-M. Baschera, “Classification of ADHD patients on the basis of independent ERP components using a machine learning system,” Nonlinear Biomedical Physics, vol. 4, no. S1, 2010.
[33] H. Oztoprak, M. Toycan, Y. K. Alp, O. Arikan, E. Dogutepe, and S. Karakas, “Machine-based learning system: Classification of ADHD and non-ADHD participants,” 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017.
[34] Y. Zhang-James, Q. Chen, R. Kuja-Halkola, P. Lichtenstein, H. Larsson, and S. V. Faraone, “Machine-Learning prediction of comorbid substance use disorders in ADHD youth using Swedish registry data,” Journal of Child Psychology and Psychiatry, vol. 61, no. 12, pp. 1370–1379, 2020.
[35] X. Peng, P. Lin, T. Zhang, and J. Wang, "Extreme learning machine-based classification of ADHD using brain structural MRI data", PloS one, vol. 8, no. 11, pp. e79476-e79476, 2013.
[36] J.W. Kim, V. Sharma and N. D. Ryan, "Predicting Methylphenidate Response in ADHD Using Machine Learning Approaches," The International Journal of Neuropsychopharmacology, vol. 18, (11), pp. pyv052-pyv052, 2015.
[37] M. Duda, R. Ma, N. Haber, and D. Wall, 2016. “Use of machine learning for behavioural distinction of autism and ADHD,” Translational Psychiatry, 6(2), pp.e732-e732, 2016.
[38] M. Duda, N., Haber, J., Daniels, and D.Wall, “Crowdsourced validation of a machine-learning classification system for autism and ADHD,” Transl Psychiatry 7, e1133, 2017. https://doi.org/10.1038/tp.2017.86
[39] G. Aggarwal, and L. Singh, “Comparisons of Speech Parameterisation Techniques for Classification of Intellectual Disability Using Machine Learning,” International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 14(2), pp.16-34, 2020.
[40] C. M. Bertoncelli, P. Vieira, E.R. Bertoncelli, D. Thummler, S. and F. Solla, "Identifying Factors Associated With Severe Intellectual Disabilities in Teenagers With Cerebral Palsy Using a Predictive Learning Model," Journal of Child Neurology, pp. 221-229, 2019.
[41] M. Cordova, K. Shada, D.V. Demeter, O. Doyle, O., Miranda-Dominguez, A. Perrone, E. Schifsky, A. Graham, E. Fombonne, B. Langhorst, J. Nigg, D.A. Fair, and E. Feczko, 2020, "Heterogeneity of executive function revealed by a functional random forest approach across ADHD and ASD," NeuroImage clinical, 2020.
[42] Y. Kou, C. Betancur, H. Xu, J.D. Buxbaum, and A. Ma'ayan, "Network- and attribute-based classifiers can prioritize genes and pathways for autism spectrum disorders and intellectual disability", American Journal of medical genetics. Part C, Seminars in medical genetics, vol. 160C, no. 2, pp. 130-142, 2012.
[43] Sewani, H., & Kashef, R. (2020). An autoencoder-based deep learning classifier for efficient diagnosis of autism. *Children*, *7*(10), 182.
[44] Kashef, R. (2021). ECNN: Enhanced Convolutional Neural Network for Efficient Diagnosis of The Autism Spectrum Disorder. *Cognitive Systems Research*.
[45] Shirazi, Z. A., de Souza, C. P., Kashef, R., & Rodrigues, F. F. (2020). Deep learning in the healthcare industry: theory and applications. In *Computational intelligence and soft computing applications in healthcare management science* (pp. 220-245). IGI Global.