ENHANCED EARLY AUTISM SCREENING: ASSESSING DOMAIN ADAPTATION WITH DISTRIBUTED FACIAL IMAGE DATASETS AND DEEP FEDERATED LEARNING

MOHAMMAD SHAFIUL ALAM[1,2], MUHAMMAD MAHBUBUR RASHID[1*]

[1]*Department of Mechatronics Engineering, IIUM, Kuala Lumpur, Malaysia*
[2]*Department of Electrical and Electronic Engineering, Northern University Bangladesh (NUB), Dhaka, Bangladesh*

**Corresponding author: mahbub@iium.edu.my*

(Received: 27 January 2024; Accepted: 30 September 2024; Published online: 10 January 2025)

ABSTRACT: This study offers a significant advancement in the area of early autism screening by offering diverse domain facial image datasets specifically designed for the detection of Autism Spectrum Disorder (ASD). It stands out as the pioneering effort to analyze two facial image datasets – Kaggle and YTUIA, using federated learning methods to adapt domain differences successfully. The federated learning scheme effectively addresses the integrity issue of sensitive medical information and guarantees a wide range of feature learning, leading to improved assessment performance across diverse datasets. By employing Xception as the backbone for federated learning, a remarkable accuracy rate of almost 90% is attained across all test sets, representing a significant enhancement of more than 30% for the different domain test sets. This work is a significant and remarkable contribution to early autism screening research due to its unique novel dataset, analytical methods, and focus on data confidentiality. This resource offers a comprehensive understanding of the challenges and opportunities in the field of ASD diagnosis, catering to both professionals and aspiring scholars.

ABSTRAK: Kajian ini menawarkan kemajuan yang ketara dalam bidang saringan awal autisme dengan menyediakan pelbagai set data imej wajah yang direka khusus untuk pengesanan Gangguan Spektrum Autisme (ASD). Kajian ini menonjol sebagai usaha perintis untuk menganalisis dua set data imej wajah – Kaggle dan YTUIA, menggunakan kaedah pembelajaran teragih untuk menyesuaikan perbezaan domain dengan jayanya. Skim pembelajaran teragih ini berkesan menangani isu integriti maklumat perubatan sensitif dan menjamin pembelajaran ciri yang meluas, yang membawa kepada prestasi penilaian yang lebih baik merentas set data yang berbeza. Dengan menggunakan Xception sebagai tunjang pembelajaran teragih, kadar ketepatan yang luar biasa hampir 90% dicapai merentas semua set ujian, mewakili peningkatan ketara lebih daripada 30% untuk set ujian domain yang berbeza. Hasil kerja ini merupakan sumbangan penting dan luar biasa dalam penyelidikan saringan awal autisme kerana set data yang unik dan baharu, kaedah analisis yang digunakan, serta tumpuan kepada kerahsiaan data. Sumber ini menawarkan pemahaman yang menyeluruh mengenai cabaran dan peluang dalam bidang diagnosis ASD, sesuai untuk para profesional dan sarjana yang berminat.

KEYWORDS: *Autism Spectrum Disorder (ASD), Artificial Intelligence, Deep Learning, Data Federation, Domain Adaptation*

1. INTRODUCTION

Rapid Autism spectrum disorder (ASD) is a neurological condition that leads to challenges in activities of daily living and communication skills for the affected person. It results from an

impairment in the development of the human brain and is characterized by a predominance of repetitive behavioral patterns and an inability to speak or communicate. The World Health Organization reports that ASD impacts 1% of babies, or one in every one hundred [1]. Unfortunately, there is presently no conclusive biomarker that can be used to diagnose ASD, nor are there any specific medications available for treating ASD[2]. However, the diagnosis is dependent on a set of standardized questionnaires and behavioral evaluations administered by experienced professionals, while the therapy is based on effective intervention and support[3]. Furthermore, current studies show that individuals diagnosed with ASD can experience substantial enhancements in cognitive and daily functioning through specific interventions and therapies before reaching the age of two. This is attributed to the ongoing development of the brain during this period, allowing for potential recovery and an increase in IQ levels up to 20 points [4].

In these circumstances, early diagnosis is essential for ASD patients to initiate appropriate intervention and cognitive therapy despite the uncertain etiology of autism in children. Preliminary study indicates that genetic factors exert a substantial influence on the likelihood of developing ASD. Research has discovered multiple genes linked to ASD, and having a family history of the illness raises the probability of its manifestation. Instead of genetic variables, researchers have investigated a limited number of prenatal and perinatal factors as potential influences on ASD. These risks include difficulties during pregnancy, exposure to specific medicines or chemicals, and maternal illnesses. In addition, scientists are studying the involvement of the immune system and gut-brain axis in the formation of ASD. The process of conventional diagnosis is quite stringent, requiring patients, together with their parents or caregivers, to undergo a series of standardized questionnaires. This can be tedious for both the patients and their guardians. This process may be biased by human expertise, age window, and parental attitudes toward their children [5].

Conversely, artificial intelligence can play a crucial function in this matter in the present era. Machine learning and deep learning, which are primary artificial intelligence techniques, can significantly automate the screening of ASD [6]. Deep learning (DL) is an algorithm created by humans that uses solid mathematical principles to identify intricate patterns in various forms of data, such as sound, images, and text, to generate precise insights and predictions. Utilizing the capacity for pattern recognition, deep learning algorithms can accurately predict autism based on specific patterns seen in brain neuroimaging data. As ASD is a neurodevelopmental disorder, these images can serve as an effective tool for identifying autism in children[7]. Deep learning models can be trained using many modalities of neuroimaging data, such as MRI and EEG thus, this model can further recognize the patterns in future samples, aiding the ASD detection process. While these brain scans show promise in detecting ASD, obtaining these samples necessitated clinical environments, sophisticated technology, and skilled physicians, all connected with significant expenses and human resources. It would be beneficial if a convenient and readily available method could be implemented for the initial assessment of ASD. Screening for ASD using face image datasets has shown significant potential as a valuable resource in numerous applications, particularly in the field of deep learning [8]. Given that the face is a direct reflection of the brain, it is reasonable to use facial traits to forecast neurological disorders such as autism [9]. Facial images offer significant advantages, including easy accessibility and the requirement for minimal preprocessing before using deep learning techniques for early screening of autism.

Deep learning relies on expansive and diverse datasets, which is crucial for training accurate models. Moreover, diverse datasets are essential for mitigating biases in machine learning, ensuring fairness and equity in applications like facial image-based pattern

recognition. While medical data sharing is essential for research, safeguarding identities remains a priority, emphasizing the delicate balance between data availability and privacy in the ever-evolving landscape of deep learning. The regulations enforced by different regulatory bodies, such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in Europe, impose additional responsibilities on data operators concerning data management and prevent the centralization of data [10]. Federated learning (FL) can serve as a solution in this context, enabling AI models to be trained utilizing decentralized datasets that are not public. Google pioneered FL, which is a decentralized machine-learning methodology that allows multiple institutions to collaborate on deep-learning projects without the need to share client data [11]. An illustrative instance of federated learning occurs when the training data is stored on the individual users' devices (nodes) instead of being sent to a central data center, which carries out computations using a local dataset to update a global model.

When utilizing several facial image datasets on separate client sites, domain differences may likely arise due to variations in imaging settings, sources, demographics, and modalities. These differences often make it challenging to create a model that is both generalizable and robust[12]. When the features of multiple datasets are not shared, accuracy is reduced across locally trained models that are isolated and applied to separate domain datasets. The federated learning system guarantees feature sharing globally and aggregates the model weight to increase accuracy among different clients; hence, it can handle multiple domain data. This study proposes an early autism screening scheme that utilizes multiple facial image datasets to train Deep CNN models. It also applies federated learning techniques for domain adaptation tasks while ensuring the security and confidentiality of individuals' crucial private information.

Table 1 presents a comparative analysis of several deep-learning models applied to a facial image dataset to detect ASD. Ying Li et al. (2023) [13] utilized an enhanced version of MobileNetV2 to identify ASD from facial images with an astounding 90.50 percent accuracy. M.F. Rabbi et al. (2023) [14] obtained an accuracy of 85 percent for ASD prediction using VGG19. H. Kang et al. (2023) [15] achieved noteworthy levels of accuracy, 85%, through the training of Xception. The AlexNet model utilized by T. M. Ghazal et al. (2023) [16] demonstrated an accuracy of 87.70 percent. All of the aforementioned studies were conducted using a single dataset, thus eliminating any potential for domain variation among the samples.

Table 1. Recent research on ASD screening using facial image dataset

Ref	Author	Algorithm	Accuracy	Dataset	Federated Learning	Domain Adaptation
[13]	Ying Li et al.	MobileNetV2	90.50	Kaggle	No	Not Required
[14]	M.F Rabbi et al.	VGG 19	85.00	Kaggle	No	Not Required
[15]	H. Kang et al.	Xception	85.00	Kaggle	No	Not Required
[16]	T. M. Ghazal et al.	AlexNet	87.70	Kaggle	No	Not Required

A new facial image dataset, YTUIA, is introduced in this study to investigate the domain adaptation efficacy over multiple source datasets. The chosen method for domain adaptation is a federated learning scheme using deep CNN in the backbone. The main contributions are

1. Introduced a new and innovative facial dataset for the purpose of comparing and investigating the impact of multiple dataset environments on ASD detection.
2. Pioneering effort in analyzing various domain datasets for domain adaptation tasks for ASD detection.

3. It innovatively employs federated learning for improved ASD screening, ensuring domain adaptation and data privacy for real-life scenarios.

The first section of the study covers an Introduction that examines contemporary research, identifies gaps in the said research, and delineates potential noteworthy contributions. The following section provides an in-depth analysis of the methodology employed in diagnosing ASD. It consists of a detailed description of the newly synthesized facial image dataset for ASD, the application of the federated learning scheme, and the domain adaptation accomplished by deploying this technique as well as maintaining data safety. Following this, the subsequent section assesses various performance indices derived from the training and testing stages of deep CNN models, as well as domain adoption, which yields more accurate predictions in scenarios involving multiple domains. Section 4 closes by providing an overview of potential future research directions and summarizing the key findings of the present study.

2. METHOD

The research study presented a suitable approach for leveraging sensitive medical datasets from different servers to improve the accuracy of ASD detection while maintaining the confidentiality of personal information. Additionally, the research endeavored ingeniously to create a new facial image dataset to examine the efficacy of domain adaptation across numerous decentralized datasets for distinct servers. The “server” is a computational node situated in different hospitals, clinics, or healthcare centers.

2.1. Datasets

2.1.1. Dataset 1 (Kaggle)

The first dataset was named "Kaggle"[1], which is compiled utilizing facial images of children with autism [17]. This dataset was readily available and easily accessible on the Kaggle repository. The dataset comprises 2D RGB images and encompasses individuals from 2 to 14 years of age, specifically focusing on toddlers between the ages of 2 and 8. The data set illustrates a male-to-female ratio of approximately 3:1, with the ratio near 1:1 for the autistic and normal control (NC) groups. Train, test, and validation sets contain 86.38%, 10.22%, and 3.41% of the sample size, respectively. Gerry Piosenka compiled the dataset exclusively from online sources, excluding demographic information, including ethnicity, clinical history, and severity of ASD.

2.1.2. Dataset 2 (TYUIA)

The second facial image dataset is called YTUIA[2] (The dataset was developed at UIA – Universiti Islam Antarabangsa). It is developed from the widely recognized video dataset named the Self-Stimulatory Behaviours dataset (SSBD) [18]. Initially, the SSBD comprised seventy-five videos. To meet the quota of one hundred videos for frame extraction, an additional fifty videos were incorporated from therapists or specialized institutions that were readily accessible on YouTube, as specific videos of the SSBD dataset were unavailable on YouTube. Following the age group, YouTube videos depicting elementary school activities were chosen as NC samples. Each frame from the videos was extracted into a distinct folder, and facial detection was performed on each frame using the MTCNN algorithm in order to eliminate frames without faces.

[1] https://github.com/mm909/Kaggle-Autism
[2] https://drive.google.com/drive/u/2/folders/1g8iyO2Q0jnWLj6w6l5Nb8vm7GogTzK0J

Following that, a rigorous preprocessing pipeline was implemented, encompassing operations such as cropping, resizing, and aligning. The "Normal control" group comprised 173 distinct members, of which 117 were male and 56 were female. The participants' age was dispersed across a range of 1 to 11 years. There were 123 individuals in the "ASD" category in the dataset; 30 were female, and 93 were male. The participants comprised an age group from 3 to 11 years. A total of 1068 samples were used in the training set, while 100 samples were used in the testing set, and a 1:1 ratio of individuals with ASD to NC was preserved.

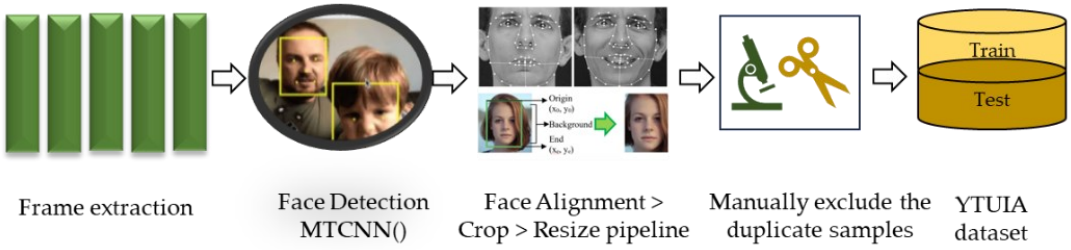


Figure 1. The process flow to develop the YTUIA dataset from video files

2.2. Domain Adaptation

Federate learning, which enables feature learning from the number of local clients remotely, has been implemented as a domain adaptation strategy to aid in the early detection of autism using facial image datasets. Domain adaptation refers to the process through which a machine learning model, trained initially on a dataset called the "source domain," is adjusted to function most effectively on a target domain dataset comprising unique characteristics. The dissimilarity between the source and target domains in facial images can be attributed to imaging apparatus and configuration variations. Transductive transfer learning is necessary to guarantee task consistency when source and target domains diverge. By facilitating the smooth transfer of features between models, this methodology guarantees the efficient adaptation of the feature space.

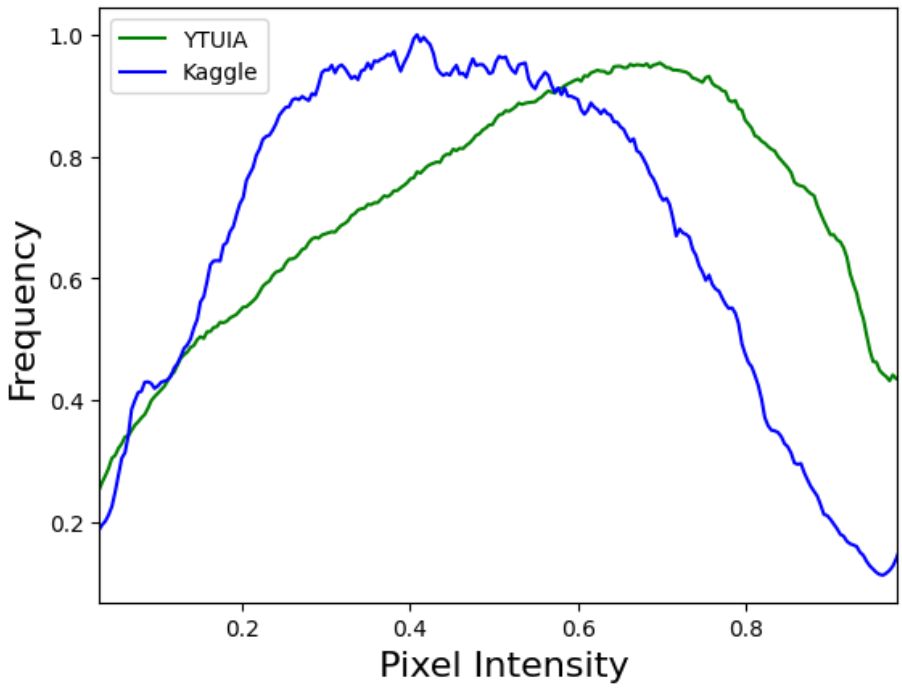


Figure 2. The domain shift in the two different facial image datasets for ASD detection

The datasets employed in this research are labeled Kaggle and YTUIA, and the corresponding test sets are denoted as T1 and T2, as described in the previous section. The normalized intensity distribution of both datasets is achieved by converting the pixel intensities of the facial images that make up the datasets, as depicted in Figure 2.

The variation in intensity distribution, indicative of a distinction in the field, leads to confusion for models when performing classification tasks. In addition, the T-statistic yields a p-value of 0.15 and a T-value of 3.10, providing strong evidence for the differentiation between the samples of the Kaggle and YTUIA datasets.

2.3. CNN Models

Convolutional Neural Networks (CNNs) have been applied to image recognition and classification in recent research. CNNs originated from investigations into the visual cortex of the brain. In recent years, CNNs have demonstrated exceptional performance on difficult visual tasks due to advancements in computing capacity and the availability of many images to train DNN. The primary function of the object classification model is to obtain features of an object, thereby enabling the binary classification of faces into two distinct classes—ASD and NC through feature learning from the facial image dataset. To derive features of both autistic and normal faces, CNN-based models that were trained using ImageNet previously can utilize the core convolutional layers. Additionally, the classification layers are adjusted to perform binary classification. MobileNetV2, ResNet50V2, and Xception models have been identified as the most effective in recent studies and are utilized here [3].

2.3.1. MobileNetV2

MobileNetV2 is an advanced algorithm that employs inverted residuals and a linear bottleneck. It is a basic module that expands for utilization in applications installed in hand phones. The algorithm has the capability to handle input with fewer dimensions, hence reducing the computational workload and memory usage while still retaining a higher accuracy level. This model employs depth-wise separable convolution, and the convolution operation is divided into two separate layers. The initial layer consists of depth-wise convolution, which is very efficient as it performs a single filtering operation. The second layer obtains new properties that are proportional to the inputs. This model greatly reduced the computational expense of typical layers of the model, where k2 acts as a multiplication factor, resulting in a savings of 8 or 9 times the computing capacity needed for 3×3 depth-wise separable convolution compared to other traditional CNNs.

2.3.2. ResNet50V2

This model consists of many bi-directional propagating residual units via identity mapping. Propagation can take place with high precision regarding classification performance between blocks. These residual mappings will greatly facilitate and enhance the training process, making it more streamlined and applicable to various scenarios. ResNet models often have a depth of over 100 layers and demonstrate outstanding performance using ImageNet.

2.3.3. Xception

This module is derived from the Inception model of Google and features a direct and modular nature. This is structured into three main components: entry, center, and exit. Each component utilizes a separable convolutional layer with activation functions Relu. The convolution layer's output has been subjected to max-pooling, with residual connections implemented at the end of each compartment. The input image with dimensions 299×299×3 is

processed by the entry flow to obtain the desired result. As a result, the output will consist of feature maps of dimensions 19×19×728. Despite undergoing nine iterations in the segment, the size of the image's features remains consistent after the middle flow. The final component produces a result with 2048 features for an input image of standard size. Finally, the prediction layer acquires the features via a fully connected (FC) layer, which is not depicted in the diagram. The ultimate layers will undergo alterations to facilitate binary categorization, as elucidated in the subsequent section.

2.4. Federated Learning

The primary concept behind employing the Federated learning scheme is to incorporate many datasets into the ASD detection process while ensuring remote operation and preserving the security and confidentiality of the data. When datasets are obtained from distinct environments, they may exhibit domain discrepancies due to variations in the equipment used and the processing pipelines employed. Therefore, it is necessary to implement an aggregated feature learning protocol to tackle this domain shift and accurately forecast the object classes for the varied domain dataset. According to Figure 3, an aggregator is required to combine the acquired features from the local client. Assume there are n diagnosis centers labeled as $C_1, C_2, \dots, C_n$, that are collecting data for ASD screening. The data samples and models are stored on local servers. Each center possesses a locally stored dataset, referred to as $\mathbf{D}l_i$. In this particular scenario, two datasets, Kaggle and YTUIA, are saved on the servers of C_1 and C_2, respectively. Each center conducts local training by training their respective datasets. For example, Center 1 trains dataset $\mathbf{D}l_1$, while Center 2 trains dataset $\mathbf{D}l_2$. Consequently, the weights obtained from local training are stored on the local server, as in Eq. (1) [19]:

$$M_i = \sum_{i,k=1}^{i,k=m} W_{i,k} F_i \tag{1}$$

where M_i is the local DL algorithm $W_{1,1}, W_{1,2}, \dots, W_{1,n}$ are the weights and $F_1, F_2, \dots, F_n$ are the features learned by m number of local models.

The process of developing a global model requires a global server and aggregator to calculate the overall weight of the DL models of the local client servers. There is no data sharing between the local clients C_1 and C_2 with the global center C_g, but the weights $W_{1,1}, W_{1,2}$ of the DL models M_1 and M_2 are shared with the global server to aggregate the weight globally and sent back to the local servers. In the end, the global aggregator makes the $FedAvg$ expressed as

$$GM = \sum_{i=1}^{i=m} h_i F_i \tag{2}$$

where GM is the averaged weight calculated by the global server considering all the features learned by the local algorithms and can be given by

$$h_i = \frac{\sum_{i=1}^{m} w_i}{n} \tag{3}$$

where n is the total number of centers in our specific case, which is 2.

2.5. Experimental Setup

The model is trained on the Kaggle platform using the TensorFlow library. Two clients for the two different domain datasets are considered two assessment centers for ASD screening. Local models‘ are trained on the two separate datasets, Kaggle and YTUIA. The test set for Kaggle will be identified as T1, and the test set for YTUIA will be named as T2. The details of the datasets are given in Table 2, where it is evident the data for the two clients is unbalanced

as the number of samples is different for the two datasets and as the datasets are collected from different persons thus, the nature of the data is not Independently and Identically Distributed (non-IID). So, the effect of the unbalanced non-IID dataset on this method can also be investigated with the domain adaptation task. Local hyperparameter was set based on the best result of the recent ablation study on these datasets [3]. The LR is set to 0.001, Adagrad is used as an optimizer, and the batch size is 32. As we work on binary classification, the loss function is set as categorical_crossentropy.

Table 2. The details of two facial image datasets

Split	Kaggle	YTUIA	Binary Class
Train set	2654	1068	0 - non-ASD 1 - ASD
Test set	280	100	
Valid set	80	-	

The deep CNN models were trained for 50 epochs and evaluated on both T1, T2, and the combined test set (T1+T2). The weights are then shared with the global server, and the model aggregator generates the global model using FedAvg. The weight of the global models is shared with each of the clients. The evaluation for T1, T2, and the combined test set was done using the global model with the federated weight from each client. To compare the evaluation performance of the federated learning approach, we commence by juxtaposing its performance with that of a conventional learning method that employs centralized and shared data to train the identical network architecture.

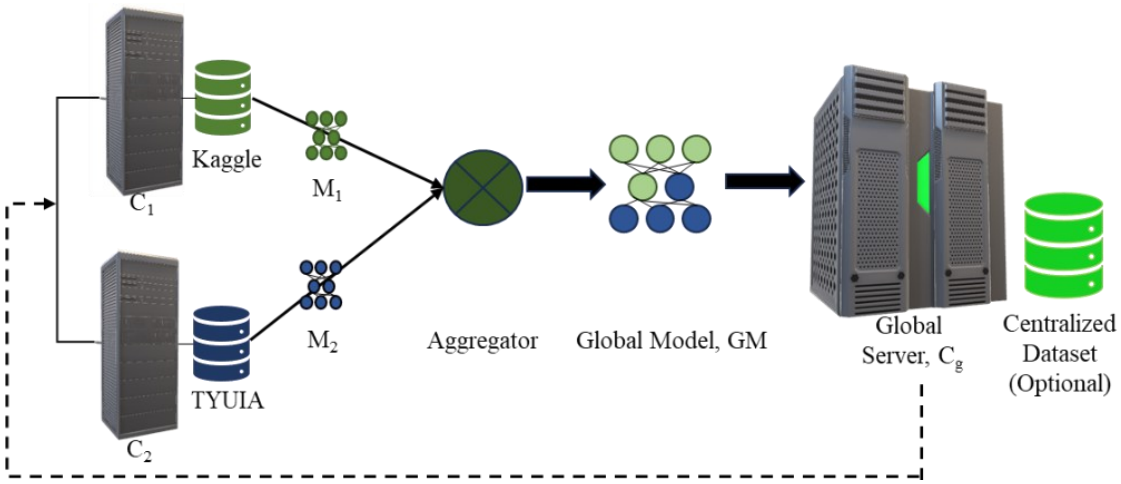


Figure 3. Process flow for federated learning for ASD detection with facial images

2.6. Evaluation Metrics

We employed several parameters to assess the performance of models. The most popular metric was the accuracy of classification, commonly referred to as 'accuracy' in this body of research. As done in earlier studies, the assessment of the model's precision also included using AUC, which stands for area under the curve. Derived from the ROC curve, this area under the curve (AUC) is a more valid measure than accuracy for illustrating the model's performance. The remaining two matrices, precision and recall, indicate the capacity to forecast the intended classes accurately. The F1-score is a metric utilized in statistical analysis of information

retrieval and binary classification systems to assess their predictive performance. The mathematical equations for accuracy, precision, and recall are as follows:

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_n + F_p} \tag{4}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{5}$$

$$Recal = \frac{T_p}{T_p + F_n} \tag{6}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

3. RESULTS AND DISCUSSION

The code was executed on the Kaggle platform, whereas the Python programming language was employed during the code development phase. Following the conclusion of the model training process, the acquired results underwent examination utilizing a variety of data analysis tools. The performance assessment in this research entailed the computation of accuracy, precision, recall, and F1-score, utilizing Equations (4) to (7).

3.1. Performance Evaluation at Local Centers

A comprehensive evaluation of three models, Xception, ResNet50V2, and MobileNetV2, trained on the Kaggle dataset, is presented in Table 3. The assessment provides critical performance matrices per the equations (4) to (7). Every model undergoes a comprehensive evaluation on two separate sets, denoted as T1 and T2, to demonstrate their efficacy under different domain scenarios. The Xception model achieved a maximum accuracy of 95% and an AUC value of 98% on T1. The ResNet50V2 and MobileNetV2 models demonstrate nearly identical evaluation performance, reaching an accuracy of 94% and 92%, respectively.

Table 3. Performance evaluation at C_1 model trained with the Kaggle dataset

Models	M_{kaggle} evaluated on T1					M_{kaggle} evaluated on T2				
	Acc	**Prec**	**Recall**	**F1**	**AUC**	**Acc**	**Prec**	**Recall**	**F1**	**AUC**
Xception	**0.95**	0.95	0.95	0.94	0.98	0.49	0.49	0.49	0.48	0.50
ResNet50V2	0.94	0.94	0.94	0.94	0.96	**0.53**	0.53	0.53	0.53	0.55
MobileNetV2	0.92	0.92	0.92	0.92	0.96	0.44	0.44	0.44	0.44	0.43

The C_1 models were evaluated on the YTUIA test set (T2), which is from a different domain, to identify the research gap. The evaluation performance has experienced a reduction of approximately 50 percent. The highest accuracy achieved was 53 percent for ResNet50V2; however, for MobileNetV2, the reported accuracy is just 44 percent, indicating a significant decrease. The result indicates a substantial disparity in performance between the two sets of tests, highlighting the need for domain adaptation when testing among various dataset samples.

Table 4 illustrates the scenario at C_2, where the local models are trained using the YTUIA dataset. As the models were evaluated on the same domain test set T2, just as C_1, their performance is quite good. ResNet50V2 obtained the highest accuracy and AUC of 95 and 97, respectively. The substantial decline in accuracy to 74% for MobileNetV2 suggests that the feature extraction by the model is unsuitable for the YTUIA dataset.

Table 4. Performance evaluation at C_2 model trained with the YTUIA dataset

Models	M_{YTUIA} evaluated on T1					M_{YTUIA} evaluated on T2				
	Acc	Prec	Recall	F1	AUC	Acc	Prec	Recall	F1	AUC
Xception	**0.57**	0.57	0.57	0.49	0.60	**0.95**	0.95	0.95	0.95	0.97
ResNet50V2	0.56	0.56	0.56	0.46	0.57	0.93	0.93	0.93	0.93	0.95
MobileNetV2	0.56	0.56	0.56	0.48	0.59	0.74	0.74	0.74	0.74	0.83

While testing on T1, a dataset obtained from Kaggle that belongs to a different domain, the assessment metrics show a significant decrease in accuracy for the Xception and ResNet50V2 models, reaching 57% and 56%, respectively.

Table 5. Performance evaluation on combined test sets locally at C_1 and C_2

Models	M_{kaggle} evaluated on T1+T2					M_{YTUIA} evaluated on T1+T2				
	Acc	Prec	Recall	F1	AUC	Acc	Prec	Recall	F1	AUC
Xception	**0.83**	0.83	0.83	0.83	0.90	0.64	0.64	0.64	0.64	0.66
ResNet50V2	0.81	0.81	0.81	0.81	0.87	**0.66**	0.66	0.66	0.66	0.69
MobileNetV2	0.82	0.82	0.82	0.82	0.86	0.60	0.60	0.60	0.60	0.64

Table 5 displays the assessment outcome conducted on the combined test set T1+T2. The results fall short of the values obtained while testing the models on the same domain test sets. The highest accuracy achieved on the combined test was 83% at C_1 while getting 66% at C_2. This issue arises due to the unbalanced number of samples in the combined test set from Kaggle and YTUIA. Due to the threefold increase in the amount of samples from Kaggle in the combined test set, the accuracy at C_1 is higher. This happens only because the models have learned features exclusively from the Kaggle dataset. For this reason, the evaluation accuracy is lower on the combined test set at C_2, where models are trained on the YTUIA dataset only.

3.2. Performance Evaluation after Federated Learning

After the training at the local servers, model weights are sent to the aggregators to update the global model, GM, as per equation (2). The global server updates the local servers with the global aggregated models. Table 6 displays the thorough performance assessment of various models, such as Xception, ResNet50V2, and MobileNetV2, utilizing global model weights on separate test sets, T1 and T2. The Xception model scored an accuracy, precision, recall, F1-score of 91%, and AUC of 99% when evaluated on T1, indicating its reliable performance. Similarly, the assessment on T2 produced slightly lower but still praiseworthy outcomes, with an accuracy of 89% and other metrics remaining consistently good.

Table 6. Performance evaluation on different test sets with global model weights

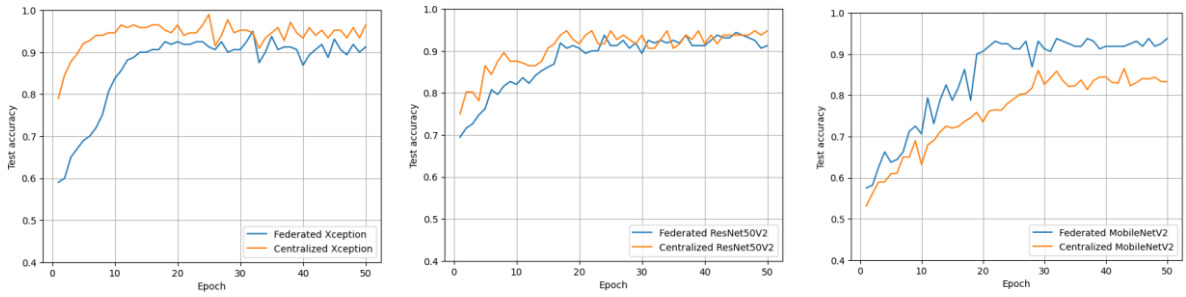
Models	$M_{Federated}$ evaluated on T1					$M_{Federated}$ evaluated on T2				
	Acc	Prec	Recall	F1	AUC	Acc	Prec	Recall	F1	AUC
Xception	**0.91**	0.91	0.91	0.91	0.99	0.89	0.89	0.89	0.89	0.98
ResNet50V2	0.90	0.90	0.90	0.90	0.96	**0.92**	0.92	0.92	0.92	0.99
MobileNetV2	0.82	0.82	0.82	0.82	0.91	0.84	0.84	0.84	0.84	0.90

The ResNet50V2 model demonstrated better performance, attaining an accuracy of 92% while evaluated on T2 than evaluation on T1, which is 90 percent. All the performance matrices consistently demonstrated the model's reliability across both test sets. Despite a slightly reduced accuracy of 82% on T1, MobileNetV2 exhibited significant precision, recall, F1-score,

and AUC values. At T2, the accuracy increased to 84 percent, confirming the model's adaptive capacity to adjust effectively, predicting various test scenarios. The variability in performance across different test sets indicates that data characteristics between T1 and T2 likely influenced the models' results, underscoring the importance of considering data heterogeneity when developing robust and generalizable models. This finding validates that the model's effectiveness is enhanced following domain adaptation, regardless of the dataset utilized for evaluation.

Moreover, when evaluating real-time unseen datasets with trained models, feature extraction and generalization are notably more adaptive after federated learning, as the model weights are averaged following training on diverse and unique datasets from different health centers.

Figure 4 shows the graphs of the proposed federated learning vs. centralized learning techniques. The results for centralized learning are the average value accumulated from both training with the Kaggle and YTUIA datasets and evaluated on the own domain dataset. The same process can also be experienced by storing the shared combined dataset in the optional global database. At first, the accuracy of the federated technique is lower than that of the centralized training data for both the Xception and ResNet50V2 models. However, after a specific number of epochs, the accuracy of the federated strategy approaches the accuracy of the centralized training data.



(a) Xception (b) ResNet50V2 (c) MobileNetV2

Figure 4. Comparison of the Centralized and Federated Learning

While the accuracy achieved with federated learning may not exceed the values obtained from training on a centralized dataset hosted on a global server, the results are consistent across all clients and test sets. The test accuracy curve obtained via federated learning surpasses shared or centralized training for MobileNetV2. The outcome from C_2, as presented in Table 4, unexpectedly demonstrates a decrease in value. This decrease has contributed to the decline of the MobileNetV2 model in the centralized training scheme.

Table 7. Performance comparison of Centralized and Federated learning on the combined set

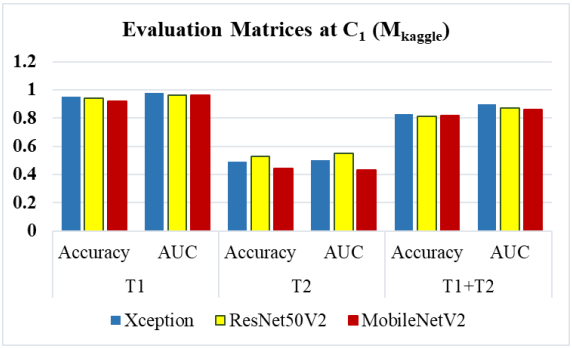
Models	**$M_{centralized}$ (avg) evaluated on T1+T2**					**$M_{Federated}$ evaluated on T1+T2**				
	Acc	**Prec**	**Recall**	**F1**	**AUC**	**Acc**	**Prec**	**Recall**	**F1**	**AUC**
Xception	0.950	0.950	0.950	0.950	0.975	0.901	0.901	0.901	0.901	0.959
ResNet50V2	0.935	0.935	0.935	0.935	0.955	0.891	0.891	0.891	0.891	0.951
MobileNetV2	0.830	0.830	0.830	0.830	0.895	0.887	0.887	0.887	0.887	0.904

Table 7 compares the evaluation matrices for the centralized and Federated training strategy when tested on the combined test set of the Kaggle and YTUIA datasets. Training models using a shared training set from all clients guarantees smooth feature learning from the training data, but it compromises the confidentiality of medical information and violates patients' privacy.

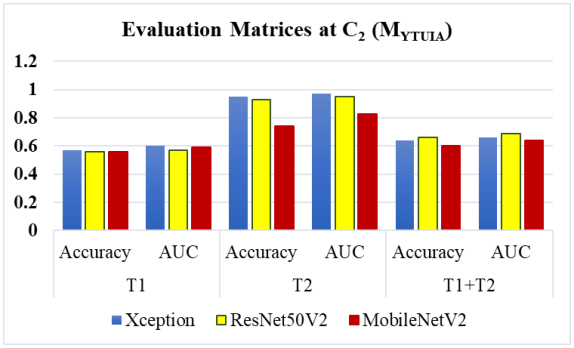
The centralized training for Xception achieves a high accuracy of 95 percent, mostly because of the open sharing of medical information. Federated learning guarantees data security and allows us to achieve up to 90 percent accuracy using the Xception model. ResNet50V2 exhibited consistent behavior throughout both schemes, while MobileNetV2 produced contrasting outcomes, which can be attributed to a lack of proper feature learning during the training phase at C_2 using the YTUIA dataset.

3.3. Discussion

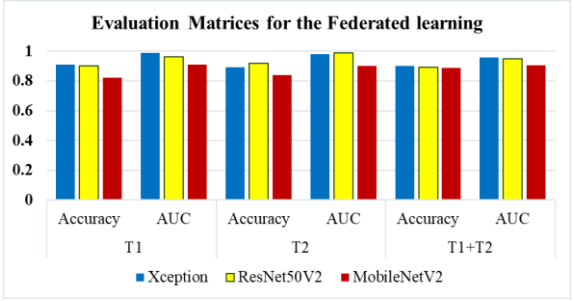
The preceding discourse constitutes a research endeavor that introduces a unique dataset of facial photos specifically designed for detecting ASD. This research also pioneers the analysis of two datasets from different domains to improve the effectiveness of domain adaptation tasks. Federated learning algorithms improve ASD screening and prioritize domain adaptability while maintaining medical information confidentiality. There are many potential challenges to employing deep federated learning for domain adaptation, and this study can contribute a few significant insights. Figure 4 displays the graph depicting the two primary performance metrics, AUC and accuracy, for different evaluation criteria. Figure 5 (a) and (b) show the performance of local models trained on a single-domain dataset.



(a) Local models trained with Kaggle



(b) Local models trained with YTUIA



(c) Federated global model

Figure 5. Evaluation matrices after testing with different test sets

However, when assessed on diverse test sets, the performance deteriorates. Observing the impact of an unbalanced training dataset at both centers, C_1 and C_2 are important. The evaluation result on the combined test set is better at C_1 due to using models trained on the Kaggle dataset. The combined test set has three times as many test samples from the Kaggle

test set as it does from YTUIA. Consequently, training using YTUIA leads to a decrease in performance on the combined dataset at C_2. Therefore, the prediction job is at stake when working at any of the local centers due to the presence of distinct features in the different samples, as opposed to the local training set. On the other hand, freely sharing personal medical information with every server is not a prudent strategy. The federated learning scheme addresses the challenges of data security through the use of certain methodologies. It also ensures diversity in feature learning by aggregating common features from local models, resulting in improved evaluation performance for test samples from different domains. Figure 5(c) shows the evaluation of the global aggregated model on all the test sets. The accuracy values for all test configurations and models are nearly identical and consistently remain near to or above 90 percent. While it may not exceed the evaluation performance of centralized training, it significantly enhances performance across all testing settings. These global models are accessible to local clients, allowing specialists from anywhere to utilize the ASD screening facility for diverse samples.

There is a lack of research on federated learning for ASD diagnosis that utilizes facial images as a tool to train the deep learning algorithm. However, this study is the first attempt to utilize federated learning to adapt a domain difference for ASD screening. Therefore, it is challenging to compare with the previous studies due to the lack of appropriate criteria that are directly relevant. Table 8 provides a concise overview of the current studies relevant to diagnosing Autism Spectrum Disorder (ASD). Only one research paper addressing facial images utilizing federated learning was in the Scopus database. However, the purpose of this paper does not involve achieving domain adaptation. In this research, a core component of federated learning is the CNN, which achieves a maximum accuracy of 60% with a 13% improvement compared to the benchmark [19]. The other two studies utilize a questionnaire-based CSV dataset. One specific research [20] employs Logistic Regression (LR), a shallow approach, while the other[21] utilizes FCNN-LSTM. Both researchers claimed a 99% accuracy rate. However, no minimum floor value for enhancement is specified. The data size is 200 for the shallow approach [20], though the subsequent study [21] does not explicitly specify the dataset size. However, Z. Fan et al. (2021) [22] employed ResNet as the backbone model for Federated learning in order to train MRI images for the purpose of ASD screening. They achieved an accuracy of 61.91%, which represents a 4% enhancement compared to the baseline value.

Table 8. Comparison with the recent research

Ref	Backbone	Algorithm	Accuracy	Improve (%)	Dataset Type	Dataset Size
Federated learning						
[20]	LR	-	99	NA	Interview Based	200
[21]	FCNN-LSTM	-	99	NA	Interview Based	NA
[19]	Own CNN	FedAVg	60	13	Facial images	2940
[22]	ResNet	FedAVg	61.92	4	Neuroimage	2131
Domain adaptation						
[23]	DCNN	MSDA	58.8	45	Neuroimage	409
[12]	MLP	FedAVg	84.9	26	Neuroimage	390
The proposed method for domain adaptation by Federated Learning						
	Xception	FedAVg	90.1	36	Facial images	3726

Regarding domain adaptation, J. Wang et al. (2020) [23] utilize neuroimaging data to train deep CNN models and employ the MSDA approach for domain adaptation. The research findings indicate an accuracy rate of 58.8%, representing a significant improvement of 45%

compared to the benchmark. The subsequent study [12] utilized a federated learning approach to attain domain similarity, resulting in an 84.9% accuracy rate, representing a 26% enhancement compared to the minimum benchmark. The sole distinction in our suggested approach lies in using facial pictures to train the deep CNNs' backbones. The utilization of Xception as the backbone framework in federated learning yielded an impressive accuracy rate of nearly 90% across all test sets, exhibiting a remarkable improvement of over 30% compared to the baseline values. Moreover, incorporating deep federated learning methods paves the way for a detailed examination of ethical concerns and for creating stronger and more secure solutions that benefit both researchers and individuals engaged in ASD screening fields.

4. CONCLUSION

This research presents a groundbreaking contribution to early autism screening by introducing a unique collection of facial image datasets specifically for detecting ASD. The innovative research of various domain datasets and the utilization of federated learning methods represent a revolutionary approach to improving domain adaptation tasks. The federated learning scheme is a reliable method that effectively tackles data security concerns and guarantees a wide range of features, resulting in enhanced assessment performance on diverse datasets. This study fills a gap in research by exploring the use of facial images for federated learning in the diagnosis of ASD. By employing Xception as the backbone of the federated learning approach, an accuracy rate of nearly 90% was achieved across all test sets. This represents a significant improvement of over 30% compared to the baseline values. However, the study does not explore explainable AI, which could offer insights into the neural network's decision-making. Future research should integrate explainable AI to improve model transparency and clinician trust. Despite this limitation, the study's distinctive dataset, methodological innovations, and ethical considerations make it a significant contribution to early autism screening research. It offers a comprehensive perspective on the challenges and opportunities in the field, serving as a valuable resource for practitioners and future researchers.

ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to the International Islamic University Malaysia for its support of M.S.A through the Tuition Fee Waiver Scheme (2021).

CODE REPOSITORY

https://github.com/gmshafiuliium/ASD-Federated-Shafi

REFERENCES

[1] Autism (2023).World Health Organization. Available: https://www.who.int/news-room/fact-sheets/ detail/autism-spectrum-disorders
[2] Rashid, M. M., & Alam, M. S. (2024). Power of Alignment: Exploring the effect of face alignment on ASD diagnosis using facial images .IIUM Engineering Journal, 25(1):317–327. https://doi.org/10.31436/iiumej.v25i1.2838
[3] Alam, M. S., Rashid, M. M., Roy, R., Faizabadi, A. R., Gupta, K. D., & Ahsan, M. M. (2022). Empirical Study of Autism Spectrum Disorder Diagnosis Using Facial Images by Improved Transfer Learning Approach. Bioengineering, 9(11):1–18. https://doi.org/10.3390/bioengineering9110710
[4] Kojovic, N., Natraj, S., Mohanty, S. P., Maillart, T., & Schaer, M. (2021). Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. Scientific Reports, 11(1):15069. https://doi.org/10.1038/s41598-021-94378-z

[5] Khodatars, M., Shoeibi, A., Sadeghi, D., Ghaasemi, N., Jafari, M., Moridian, P., Khadem, A., Alizadehsani, R., Zare, A., Kong, Y., Khosravi, A., Nahavandi, S., Hussain, S., Acharya, U. R., & Berk, M. (2021). Deep learning for neuroimaging-based diagnosis and rehabilitation of Autism Spectrum Disorder: A review. Computers in Biology and Medicine, 139:104949. https://doi.org/10.1016/j.compbiomed.2021.104949

[6] Mohammad Shafiul Alam, Zabina Tasneem, Sher Afghan Khan, & Muhammad Mahbubur Rashid. (2023). Effect of Different Modalities of Facial Images on ASD Diagnosis using Deep Learning-Based Neural Network. Journal of Advanced Research in Applied Sciences and Engineering Technology, 32(3):59–74. https://doi.org/10.37934/araset.32.3.5974

[7] Uddin, M. Z., Shahriar, M. A., Mahamood, M. N., Alnajjar, F., Pramanik, M. I., & Ahad, M. A. R. (2024). Deep learning with image-based autism spectrum disorder analysis: A systematic review. Engineering Applications of Artificial Intelligence, 127:107185. https://doi.org/10.1016/j.engappai.2023.107185

[8] Alam, M. S., Rashid, M. M., Faizabadi, A. R., Mohd Zaki, H. F., Alam, T. E., Ali, M. S., Gupta, K. D., & Ahsan, M. M. (2023). Efficient Deep Learning-Based Data-Centric Approach for Autism Spectrum Disorder Diagnosis from Facial Images Using Explainable AI. Technologies, 11(5):115. https://doi.org/10.3390/technologies11050115

[9] Vuilleumier, P., & Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. Neuropsychologia, 45(1): 174–194. https://doi.org/10.1016/j.neuropsychologia.2006.06.003

[10] Vizitiu, A., Nita, C. I., Puiu, A., Suciu, C., & Itu, L. M. (2019). Towards Privacy-Preserving Deep Learning based Medical Imaging Applications. 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 1–6. https://doi.org/10.1109/MeMeA.2019.8802193

[11] Feki, I., Ammar, S., Kessentini, Y., & Muhammad, K. (2021). Federated learning for COVID-19 screening from Chest X-ray images. Applied Soft Computing, 106:107330. https://doi.org/10.1016/j.asoc.2021.107330

[12] Li, X., Gu, Y., Dvornek, N., Staib, L. H., Ventola, P., & Duncan, J. S. (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Medical Image Analysis, 65:101765. https://doi.org/10.1016/j.media.2020.101765

[13] Li, Y., Huang, W.-C., & Song, P.-H. (2023). A face image classification method of autistic children based on the two-phase transfer learning. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1226470

[14] Rabbi, M. F., Zohra, F. T., Hossain, F., Akhi, N. N., Khan, S., Mahbub, K., & Biswas, M. (2023). Autism Spectrum Disorder Detection Using Transfer Learning with VGG 19, Inception V3 and DenseNet 201, Communications in Computer and Information Science, Springer, Cham, 1704:190–204. https://doi.org/10.1007/978-3-031-23599-3_14

[15] Kang, H., Yang, M., Kim, G.-H., Lee, T.-S., & Park, S. (2023). DeepASD: Facial Image Analysis for Autism Spectrum Diagnosis via Explainable Artificial Intelligence. 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), 2023:625–630. https://doi.org/10.1109/ICUFN57995.2023.10200203

[16] M. Ghazal, T., Munir, S., Abbas, S., Athar, A., Alrababah, H., & Adnan Khan, M. (2023). Early Detection of Autism in Children Using Transfer Learning. Intelligent Automation & Soft Computing, 36(1):11–22. https://doi.org/10.32604/iasc.2023.030125

[17] Musser, M. (2020). Detecting Autism Spectrum Disorder in Children With Computer Vision. Towards Data Science. Available: https://github.com/mm909/Kaggle-Autism

[18] Rajagopalan, S. S., & Goecke, R. (2014). Detecting self-stimulatory behaviours for autism diagnosis. 2014 IEEE International Conference on Image Processing (ICIP), 2014:1470–1474. https://doi.org/10.1109/ICIP.2014.7025294

[19] Shamseddine, H., Otoum, S., & Mourad, A. (2022). On the Feasibility of Federated Learning for Neurodevelopmental Disorders: ASD Detection Use-Case. GLOBECOM 2022 - 2022 IEEE Global Communications Conference, 1121–1127. https://doi.org/10.1109/GLOBECOM48099.2022.10001248

[20] Farooq, M. S., Tehseen, R., Sabir, M., & Atal, Z. (2023). Detection of autism spectrum disorder

(ASD) in children and adults using machine learning. Scientific Reports, 13(1): 9605. https://doi.org/10.1038/s41598-023-35910-1

[21] Lakhan, A., Mohammed, M. A., Abdulkareem, K. H., Hamouda, H., & Alyahya, S. (2023). Autism Spectrum Disorder detection framework for children based on federated learning integrated CNN-LSTM. Computers in Biology and Medicine, 166:107539. https://doi.org/10.1016/j.compbiomed.2023.107539

[22] Fan, Z., Su, J., Gao, K., Hu, D., & Zeng, L.-L. (2021). A Federated Deep Learning Framework for 3D Brain MRI Images. 2021 International Joint Conference on Neural Networks (IJCNN), 1–6. https://doi.org/10.1109/IJCNN52387.2021.9534376

[23] Wang, J., Zhang, L., Wang, Q., Chen, L., Shi, J., Chen, X., Li, Z., & Shen, D. (2020). Multi-Class ASD Classification Based on Functional Connectivity and Functional Correlation Tensor via Multi-Source Domain Adaptation and Multi-View Sparse Representation. IEEE Transactions on Medical Imaging, 39(10): 3137–3147. https://doi.org/10.1109/TMI.2020.2987817