DeepInsight: Multi-Task Multi-Scale Deep Learning for Mental Disorder Diagnosis

Mingyu Ding[1]
d130143597@163.com
Yuqi Huo[2]
bnhony@163.com
Jun Hu[2]
junhu@ruc.edu.cn
Zhiwu Lu[1]
luzhiwu@ruc.edu.cn

[1] School of Information
Renmin University of China
Beijing, 100872, China

[2] Beijing Key Laboratory
of Big Data Management
and Analysis Methods
Beijing, 100872, China

Abstract

We propose a novel deep learning approach, called DeepInsight, to quick diagnosis of autism spectrum disorder (ASD) and major depressive disorder (MDD). Our approach is motivated by recent advances in artificial intelligence (AI) for healthcare. In particular, researchers have found distinct differences between facial characteristics of children with ASD and those of typically developing children. Based upon these findings, we choose to explore deep learning in extracting discriminative facial features. However, for ASD diagnosis, the labelled data are far from sufficient for training a deep learning model. Therefore, by considering the two typical mental disorders (i.e. ASD and MDD) together, we develop a multi-task deep learning model to augment the labelled data for each diagnosis task. Moreover, we also induce multi-scale combination into the proposed model to learn more discriminative facial features. Experimental results demonstrate the effectiveness and efficiency of our approach to mental disorder diagnosis.

1 Introduction

Artificial intelligence (AI) has been widely leveraged in many healthcare applications such as skin cancer classification [7], congenital cataract management [23], and personalized nutrition [37], due to the latest advances in machine learning (especially deep learning). As one of the most challenging problems in healthcare, autism spectrum disorder (ASD) diagnosis has also draw much attention [2, 12, 15, 22, 34] from both psychiatry and AI.

The focus of this paper is also ASD diagnosis, which is inspired by [2, 3]. In these closely related works, facial characteristics of children with ASD were shown to have distinct differences from those of typically developing (TD) children. For example, [2, 3] have reported the following observations: 1) children with ASD have a broader upper face, including wider eyes; 2) children with ASD have a shorter middle region of the face, including the cheeks and nose. According to these findings, we choose to extract discriminative facial features for ASD diagnosis by employing deep learning [17] as basic AI tool.

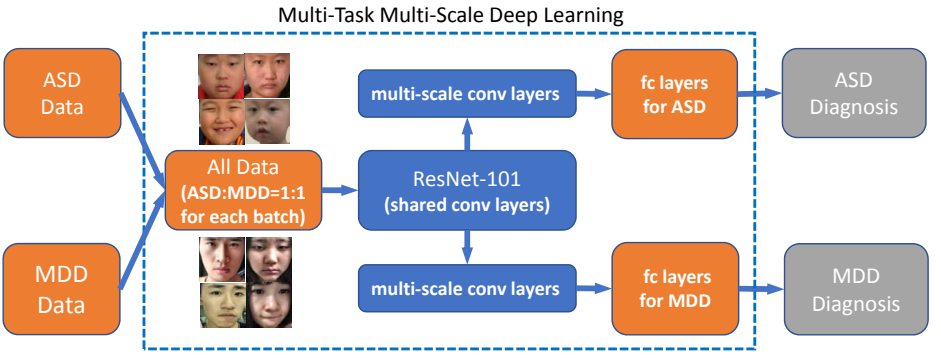

Figure 1: The flowchart of our DeepInsight approach to quick diagnosis of the two metal disorders (i.e. ASD and MDD). The conv and fc layers in the proposed model denote the convolutional and fully-connected layers, respectively.

Although deep learning has been shown to yield exciting results in many fields [10, 26, 27, 35, 38], it still has a limitation on model training when applied to healthcare problems. That is, the ground-truth labels of medical data are very expensive to access, and we are generally provided with a small labelled set for model training. It is well-known that the scarcity of training data tends to cause the overfitting of a deep learning model [18, 19]. Hence, when leveraging deep learning in ASD diagnosis, our focus is how to overcome the overfitting issue during model training.

In this paper, by considering another typical mental disorder, i.e., major depressive disorder (MDD), together with ASD, we develop a multi-task deep learning model to augment the labelled data for each diagnosis task. As compared to the traditional single-task deep learning that trains a convolutional neural network (CNN) model [13, 16, 30, 32] for each diagnosis task, multi-task deep learning trains only a single CNN model for multiple diagnosis tasks, similar to [40, 42]. Note that the distinct advantage of multi-task deep learning is that the labelled data of multiple diagnosis tasks can be shared to train a more robust CNN model, which is crucial for leveraging deep learning in mental disorder diagnosis. Although multi-task learning has been successfully applied to many medical problems [5, 31, 34] in the literature, rare attention has been paid to multi-task diagnosis of *more than one disorders*.

As one of the most advancing CNN models, ResNet-101 [40] is used to form the shared convolutional (conv) layers in our multi-task deep learning model. However, the shared conv layers can only extract the shared facial features for the diagnosis of ASD and MDD. To ensure that each diagnosis task has its own conv layers for task-aware feature learning, we thus induce multi-scale combination [20, 25] into our multi-task deep learning model, which is denoted with multi-scale conv layers in Figure 1. Moreover, by adopting the pre-training and finetuning strategies, we develop a robust algorithm to train the proposed multi-task multi-scale deep learning model. Note that the proposed model can not only learn shared facial features but also learn task-aware facial features for the diagnosis of ASD and MDD, by taking both multi-task deep learning and multi-scale combination into consideration.

To evaluate the performance of our DeepInsight approach, we conduct extensive experiments on two face datasets (denoted as ASD-Face and MDD-Face), which are collected from a local psychiatric hospital and also from the web (e.g., the we-media on the ASD and

MDD topics, and documentary films about ASD and MDD). Experimental results show the effectiveness of our approach in the two diagnosis tasks. This means that both shared and task-aware facial features are indeed crucial for the diagnosis of ASD and MDD. In addition, our results also show that our approach (taking less than 1 second per-subject) is much more efficient than the clinical diagnosis of ASD and MDD (taking more than 0.5 hour per-subject) based on the behaviors of the subjects.

Our main contributions are: (1) We have proposed a multi-task deep learning approach to the diagnosis of *more than one mental disorders*, which has been rarely considered in the literature. (2) We have made the first contribution to learning *both shared and task-aware* deep features for multi-task medical diagnosis, to the best of our knowledge. (3) We have developed a robust algorithm to train the proposed multi-task deep learning model.

2 Related Work

ASD Diagnosis. ASD is a neurodevelopmental disorder defined in DSM-5. Children with ASD must present two types of symptoms: 1) deficits in social communication and social interaction; 2) restricted, repetitive patterns of behavior, interests or activities. Among various psychological assessment tools, the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) are considered the "gold standards" for assessing autistic children. However, the clinical diagnosis of ASD with these measurements is subject to the expertise of psychiatrists, and the whole procedure may continue several months (including multiple times of clinical diagnosis). To overcome these limitations in clinical diagnosis, many AI methods have been developed for ASD diagnosis. Specifically, children with ASD are recognized from typically developing children by machine learning with various types of medical data including genes [15], magnetic resonance imaging (MRI) of brain [12, 34], and eye movements [8, 22]. In this paper, according to the interesting findings in [2, 8], we propose a deep learning model to extract discriminative facial features for ASD diagnosis. Since the face pictures of children can be obtained at significantly less cost (of time and money) than genes, brain MRI, and eye movements, our model is expected to have a wider use in real-world applications.

MDD Diagnosis. MDD is a mental disorder characterized by at least two weeks of low mood. It often comes along with low self-esteem, loss of interest in normally enjoyable activities, low energy, and pain without a clear cause. The most widely used criteria for diagnosing depressive conditions can be found in DSM-IV-TR and ICD-10. Based on these assessment measurements, the clinical diagnosis of MDD is subject to the expertise of psychiatrists (or psychologists), which is similar to the clinical diagnosis of ASD. Therefore, biomarkers of MDD have been explored to provide an objective method for clinical diagnosis. There exist several potential biomarkers, including brain-derived neurotrophic factor [33] and various functional MRI techniques [11]. However, no biological assessments can confirm major depression, and more effective biomarkers are needed for MDD diagnosis.

Multi-Task Learning in Healthcare. Since multi-task learning [40, 42] can greatly alleviate the scarcity of training data, it has been successfully applied to many medical problems [5, 31, 34]. In [34], multi-task ASD diagnosis was performed across multiple medical imaging centers, i.e., a single task refers to the ASD diagnosis for one medical imaging center. In [31], the neuroimaging data was first grouped into multiple subclasses by a clustering method, and then an effective approach to Alzheimer's disease diagnosis was proposed based on multi-task learning across multiple subclasses. In [5], a family of multi-task learning algo-

rithms were developed for collaborative computer aided diagnosis over multiple clinically-related datasets of medical images. It can be seen that these multi-task learning methods have considered *only one medical disorder*. In contrast, we focus on multi-task diagnosis of more than one medical disorders. This also means that the conventional multi-task learning methods are not suitable for our application scenario.

Deep Learning in Healthcare. We face great challenges when applying deep learning is applied to healthcare problems. Specifically, the ground-truth labels of medical data are very expensive to access, and a small labelled set is generally provided for training deep learning models. Due to the scarcity of training data, the overfitting issue tends to degrade the performance of deep learning [18, 19]. Hence, when leveraging deep learning in the diagnosis of ASD and MDD, our focus is how to overcome the overfitting issue during model training. In this paper, we adopt the multi-task learning strategy to alleviate the scarcity of training data, and also develop a robust algorithm for model training.

3 The Proposed Model

3.1 Face Preprocessing

Similar to previous work on face recognition [1, 4, 6, 9, 21], we preprocess the original large face pictures to obtain standard faces, before training deep learning models. In this paper, face preprocessing includes:

Face Detection. We first detect a single face or multiple faces from each original large picture with FaceNet [29]. For each detected face, we further detect 68 facial keypoints.

Face Alignment & Cropping. Based on the detected 68 keypoints, we align each face using a 2D affine transformation and then crop it to the size 256*256 pixels.

Data Augmentation. The output of face alignment & cropping is of the size 256*256 pixels, but the input size of ResNet-101 is 224*224 pixels. Instead of random cropping used in Caffe, we adopt the following data augmentation method: we first crop each original face of 256*256 pixels at five positions, denoted by the upper-left corners (0, 0), (0, 32), (16, 16), (32, 0) and (32, 32), to generate 5 new faces of the size 224*224 pixels and then horizontally flip the five cropped faces to double the face number.

The data augmentation step has three advantages: 1) the effect of head pose on model training can be suppressed to some extent; 2) the deep learning model can be trained with more data to avoid the overfitting problem; 3) the 10 new cropped faces can be used to compute a classification probability for each face from the test set.

3.2 Network Architecture

For the diagnosis of ASD and MDD, we design a multi-task multi-scale deep learning model based on a Caffe implementation of ResNet-101 [40], which is a typical CNN model. The network architecture of our deep learning model is illustrated in Figure 2. In this deep learning model, there are three groups of convolutional layers: (1) the first group of shared convolutional layers for the two diagnosis tasks, called ResNet-101; (2) the second group of multi-scale convolutional layers for ASD diagnosis, called conv_ASD; (3) the third group of multi-scale convolutional layers for MDD diagnosis, called conv_MDD.

In this paper, we regard each group of convolutional layers as a subnetwork in our deep learning model. The details of the three subnetworks are given as follows:

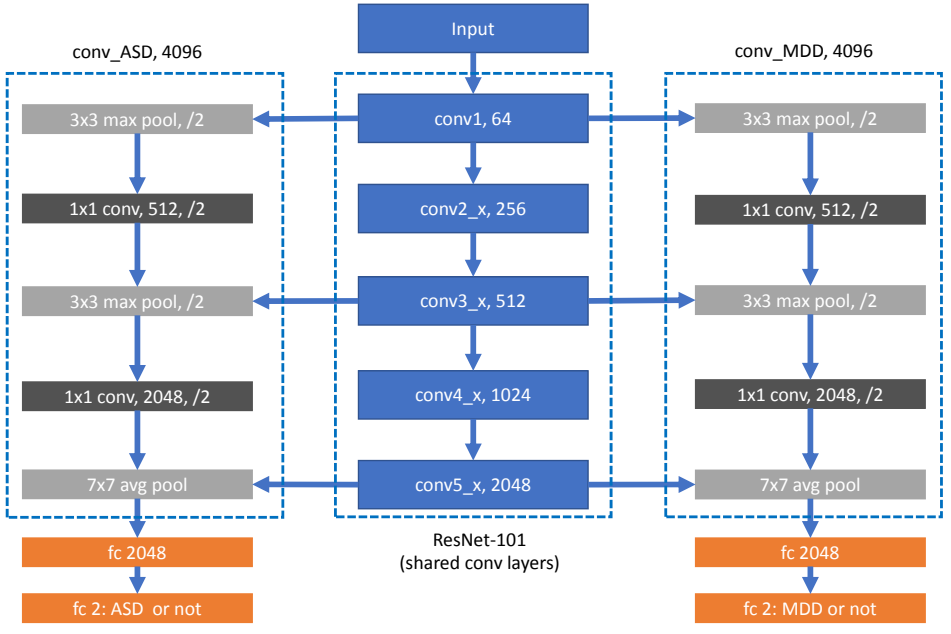

Figure 2: The network architecture of our multi-task multi-scale deep learning model.

ResNet-101. This subnetwork is generally inherited from the original ResNet-101 model, which consists of five subgroups of convolutional layers. In this paper, we modify the original ResNet-101 in two aspects: (1) The fully-connected layers of ResNet-101 are not included this subnetwork; (2) In the input layer, half of each batch contains samples from the ASD-Face dataset, and the other half contains samples from MDD-Face. In our model, this subnetwork is used to extract shared discriminative facial features for the diagnosis of ASD and MDD, which is exactly consistent with the goal of multi-task learning.

conv_ASD. The design of the conv_ASD subnetwork is inspired by multi-scale combination [20, 25] in CNN models. In the ResNet-101 subnetwork, there are five subgroups of convolutional layers and each subgroup contains one or multiple convolutional layers. For the trade-off between efficiency and effectiveness, we only integrate three layers of ResNet-101 (i.e. the last layer of the first, third, and fifth subgroups of convolutional layers), as shown in Figure 2. This subnetwork consists of two 3×3 max pooling layers, two 1×1 convolutional layers, and one average pooling layer. In our model, the conv_ASD subnetwork is used to extract task-aware discriminative facial features for ASD diagnosis.

conv_MDD. The conv_MDD subnetwork is designed similarly to conv_ASD, which also contains two 3×3 max pooling layers, two 1×1 convolutional layers, and one average pooling layer. In our model, the conv_MDD subnetwork is used to extract task-aware discriminative facial features for MDD diagnosis.

At the end of conv_ASD (or conv_MDD), two fully-connected layers are added for the diagnosis of ASD (or MDD). This also enables us to train our multi-task multi-scale deep learning model in an end-to-end manner.

3.3 Model Training

In this subsection, we develop a robust algorithm to train our multi-task multi-scale deep learning model (see Figure 2), by adopting the pre-training and finetuning strategies.

We first define the loss of our model as follows. Let x_0 and x_1 be the output of the last fully-connected layer at the end of conv_ASD, and $l_{\text{ASD}} \in \{0,1\}$ be the label of the current sample in the ASD-Face dataset. The softmax loss for ASD diagnosis is defined as:

$$\hat{x}_i = x_i - \max(x_0, x_1), i = 0, 1 \tag{1}$$

$$p_{\text{ASD}}(i) = e^{\hat{x}_i}/(e^{\hat{x}_0} + e^{\hat{x}_1}), i = 0, 1 \tag{2}$$

$$\text{Loss}_{\text{ASD}} = -\log p_{\text{ASD}}(l_{\text{ASD}}) \tag{3}$$

Similarly, let y_0 and y_1 be the output of the last fully-connected layer at the end of conv_MDD, and $l_{\text{MDD}} \in \{0,1\}$ be the label of the current sample in the MDD-Face dataset. The softmax loss for MDD diagnosis is defined as:

$$\hat{y}_i = y_i - \max(y_0, y_1), i = 0, 1 \tag{4}$$

$$p_{\text{MDD}}(i) = e^{\hat{y}_i}/(e^{\hat{y}_0} + e^{\hat{y}_1}), i = 0, 1 \tag{5}$$

$$\text{Loss}_{\text{MDD}} = -\log p_{\text{MDD}}(l_{\text{MDD}}) \tag{6}$$

The loss of our multi-task deep learning model is given by:

$$\text{Loss} = \lambda_{\text{ASD}}\text{Loss}_{\text{ASD}} + \lambda_{\text{MDD}}\text{Loss}_{\text{MDD}} \tag{7}$$

where λ_{ASD} and λ_{MDD} are the weights for the two diagnosis tasks. In this paper, we assume that the two diagnosis tasks have the same importance, and set λ_{ASD} and λ_{MDD} equal to 1 for all experiments. Note that optimizing the above loss leads to an end-to-end training process for our Multi-Task Multi-Scale ResNet model.

Since the fusion of ASD-Face and MDD-Face is still "small" for training a deep learning model, we explore CASIA-WebFace [36] and CK+ [24] as outside data for model training. In particular, CASIA-WebFace is a large-scale face dataset of 10,575 subjects and 494,414 face pictures, and CK+ is a facial expression dataset of 2,977 face pictures from eight emotion categories (i.e., neutral, anger, contempt, disgust, fear, happy, sadness, and surprise). In this paper, the CASIA-WebFace and CK+ datasets are used for model pre-training and finetuning, respectively. Our robust algorithm for model training is outlined as follows:

- **Step 1:** Pre-train a basic ResNet-101 model with the CASIA-WebFace dataset;
- **Step 2:** Finetune the pre-trained ResNet-101 with the CK+ emotion dataset;
- **Step 3:** Initialize the Multi-Task ResNet model (which simplifies both conv_ASD and conv_MDD to average pooling) using the finetuned ResNet-101 model.
- **Step 4:** Finetune all the parameters of Multi-Task ResNet with the two training sets of ASD-Face and MDD-Face.
- **Step 5:** Initialize Multi-Task Multi-Scale ResNet using finetuned Multi-Task ResNet.
- **Step 6:** Finetune all the parameters of Multi-Task Multi-Scale ResNet with the two training sets of ASD-Face and MDD-Face.

Dataset	Training Set (positives/negatives)	Test Set (positives/negatives)
ASD-Face	253/320	51/69
MDD-Face	240/240	60/60

Table 1: The characteristics of the two face datasets.

3.4 Test Process

Once our Multi-Task Multi-Scale ResNet model has been well trained, we can evaluate its performance on the two test sets of ASD-Face and MDD-Face as follows. For each pair of test faces $(I_{\text{ASD}}, I_{\text{MDD}})$, we preprocess them using the face preprocessing method and obtain 10 pairs of augmented test faces $(\hat{I}^{(i)}_{\text{ASD}}, \hat{I}^{(i)}_{\text{MDD}})(i = 1, ..., 10)$. We then input each pair of augmented test faces $(\hat{I}^{(i)}_{\text{ASD}}, \hat{I}^{(i)}_{\text{MDD}})$ into our multi-task deep learning model, and predict their labels as: $(l^{(i)}_{\text{ASD}}, l^{(i)}_{\text{MDD}})$, where $l^{(i)}_{\text{ASD}} \in \{0,1\}$ and $l^{(i)}_{\text{MDD}} \in \{0,1\}$. Given that ASD is denoted by $l^{(i)}_{\text{ASD}} = 1$ and MDD is denoted by $l^{(i)}_{\text{MDD}} = 1$, we predict the label l_{ASD} of test face I_{ASD} and the label l_{MDD} of test face I_{MDD} as:

$$l_{\text{ASD}} = \begin{cases} 1 & , \ \sum_{i=1}^{10} l^{(i)}_{\text{ASD}} > 3 \\ 0 & , \ \text{otherwise} \end{cases} \tag{8}$$

$$l_{\text{MDD}} = \begin{cases} 1 & , \ \sum_{i=1}^{10} l^{(i)}_{\text{MDD}} > 3 \\ 0 & , \ \text{otherwise} \end{cases} \tag{9}$$

where the threshold of 3 is empirically selected by taking the trade-off between the accuracies of recognition of positives and negatives in the two diagnosis tasks. This threshold can also be selected by cross-validation on the training set. Although the training process of our Multi-Task Multi-Scale ResNet model is time-consuming, the above test process is very efficient since only forward computation is used for test with the trained model.

4 Experimental Evaluation

4.1 Data Collection

For performance evaluation, we construct two datasets[1]: ASD-Face of 693 face pictures, and MDD-Face of 600 face pictures. The characteristics of the two datasets are given in Table 1. To make the two datasets as large as possible, we have collected the face pictures not only from a prestigious local psychiatric hospital but also from the web (e.g., the we-media on the ASD and MDD topics, and documentary films about ASD and MDD). Note that the datasets collected in this way would inevitably have noise. As a remedy, we have made great effort on quality assurance during data collection, i.e., each case has been checked by at least two psychiatrists. In addition, our DeepInsight project has been confirmed by the Ethics Committee of the local hospital.

[1] https://github.com/anonymous04321/Face-Datasets-of-Mental-Disorders

Models	ASD					MDD				
	ACC	SEN	SPE	PPV	NPV	ACC	SEN	SPE	PPV	NPV
EigenFace+SVM	61.7	56.7	62.5	66.7	60.8	63.3	60.8	64.8	71.7	61.7
FisherFace+LDA	69.2	58.7	84.1	79.1	75.1	68.3	62.6	78.5	79.8	69.5
ResNet-101	85.8	80.4	87.3	87.0	83.3	83.3	81.8	82.3	86.1	81.3
Multi-Scale ResNet	87.5	82.4	91.3	87.5	87.5	85.8	86.4	83.3	87.7	83.6
Multi-Task ResNet	88.3	**88.2**	88.4	87.9	**91.0**	86.7	89.4	83.3	86.8	86.5
Full CNN Model	**90.0**	84.3	**94.2**	**91.5**	89.0	**89.2**	**92.4**	**85.2**	**88.4**	**90.2**

Table 2: Comparison among different diagnosis models for ASD and MDD diagnosis.

4.2 Experimental Settings

We evaluate the performance of ASD diagnosis (or MDD diagnosis) on the test set of ASD-Face (or MDD-Face). As in previous work on healthcare problems, five measures are used for performance evaluation: accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), and negative predictive value (NPV). Given the number of true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN), the five measures are defined as follows:

$$\mathrm{ACC} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}) \tag{10}$$

$$\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}),\ \mathrm{SPE} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP}) \tag{11}$$

$$\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}),\ \mathrm{NPV} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FN}) \tag{12}$$

In the following experiments, we train our Multi-Task Multi-Scale ResNet model in an end-to-end manner using back-propagation [28] and stochastic gradient descent [41]. We randomly initialize the new fully-connected layers (at the end of conv_ASD and conv_MDD) of our model by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01, and initialize the bias to 0. For all the new convolutional layers (i.e., conv_ASD and conv_MDD) of our model, we adopt the Xavier initialization. All the other layers (i.e., the shared conv layers) of our model are initialized by the original ResNet-101 model trained with the CASIA-WebFace and CK+ datasets. A learning rate of 0.0001 is set for the first 1,000 mini-batches, and reduced to 0.1 times with a step size of 1,000. The maximum number of iterations is set to 3,000. A momentum of 0.9 and a weight decay of 0.01 are also set for model training. In particular, we train the single-task models with 2 GPUs (batch size = 10), and train the multi-task models with 4 GPUs (batch size = 5). Our implementation is developed using Python based on the Caffe framework.

4.3 Diagnosis Results

Comparative Evaluation. We select six diagnosis models for performance evaluation: 1) **EigenFace+SVM** – the SVM classifier with the facial features extracted by EigenFace [39]; 2) **FisherFace+LDA** – the linear discriminant analysis (LDA) classifier with the facial features extracted by FisherFace [14]; 3) **ResNet-101** – the original ResNet-101 model; 4) **Multi-Scale ResNet** – the original ResNet-101 followed by conv_ASD (or conv_MDD); 5) **Multi-Task ResNet** – the degraded Multi-Task Multi-Scale ResNet by simplifying both conv_ASD and conv_MDD to average pooling; 6) **Full CNN Model** – our Multi-Task Multi-Scale ResNet model illustrated in Figure 2. The first four diagnosis models are all used for

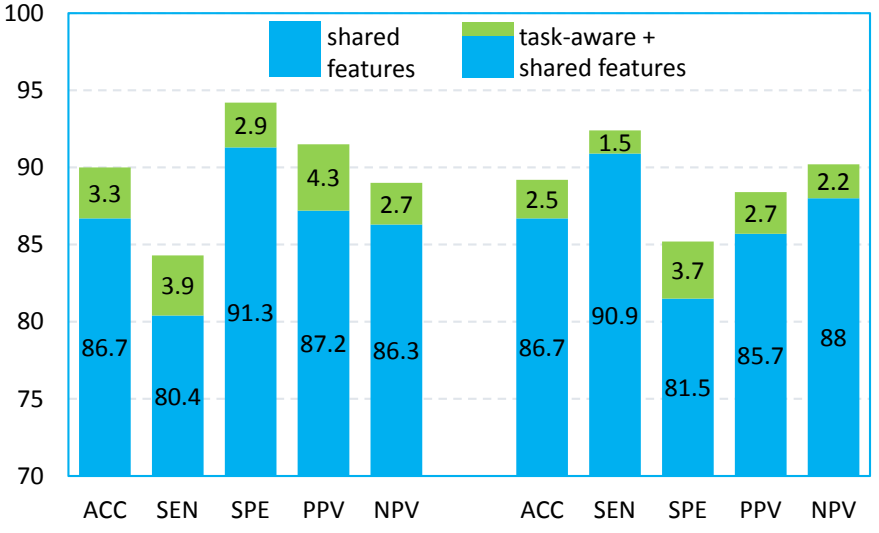

Figure 3: Comparative results of ASD diagnosis (*left*) and MDD diagnosis (*right*) using the shared features/task-aware+shared features learned by our DeepInsight model.

the two diagnosis tasks independently. The four deep learning models among these diagnosis models are all initialized using the outside data CASIA-WebFace and CK+ before model training. Note that the multi-task learning problem defined in this paper (*both inputs and outputs are distinct*) cannot be solved by the conventional multi-task learning models that typically take the same inputs or the same outputs. Therefore, we do not include these multi-task learning models in our performance evaluation.

Table 2 shows the results obtained by the above six models for the diagnosis of ASD and MDD, respectively. It can be seen that: 1) By overall evaluation, our Multi-Task Multi-Scale ResNet model performs the best among all the six models. The superior performance of our model is mainly due to the fact that it can extract both shared and task-aware facial features for mental disorder diagnosis. This is also supported by the results given by Figure 3, where the shared features yield dominant results and the task-aware features lead to further improvements. 2) The gains achieved by Multi-Scale ResNet over ResNet-101 demonstrate the effectiveness of multi-scale combination for the two diagnosis tasks. This means that Multi-Scale ResNet can extract more discriminative facial features for mental disorder diagnosis than ResNet-101. 3) The gains achieved by Multi-Task Multi-Scale ResNet over Multi-Scale ResNet demonstrate the effectiveness of multi-task learning for the two diagnosis tasks. That is, the shared conv layers can help to alleviate the scarcity of training data. 4) Multi-Task ResNet is shown to yield promising results (especially for ASD diagnosis). This provides extra evidence that multi-task learning is effective for mental disorder diagnosis. 5) As expected, all the four deep learning models yield significant improvements over the conventional classifiers using hand-craft features.

CNN Alternatives. In this paper, we employ ResNet-101 as a basic CNN model to develop our multi-task multi-scale model for ASD and MDD diagnosis. In fact, any CNN model can be used to design our network architecture. We compare ResNet-101 to VGG-16 [30] and Inception-ResNet v1 [32] by applying them to ASD and MDD diagnosis separately.

Models	ASD					MDD				
	ACC	SEN	SPE	PPV	NPV	ACC	SEN	SPE	PPV	NPV
ResNet-101	85.8	80.4	**87.3**	**87.0**	83.3	83.3	81.8	**82.3**	**86.1**	81.3
VGG-16 (VGGFace)	85.0	89.4	79.6	84.3	86.0	84.2	87.9	79.6	84.1	84.3
Inception-ResNet v1	**86.7**	**90.2**	84.1	80.7	**92.1**	**84.8**	**88.4**	81.5	85.6	**85.3**

Table 3: Comparative results obtained by different CNN models used for ASD and MDD diagnosis independently.

The same experimental setting is adopted for each CNN model. The comparative results are shown in Table 3. We have the following observations: (1) The three CNN models generally yield comparable results in the two tasks of ASD and MDD diagnosis. (2) Both ResNet-101 and Inception-ResNet v1 achieve slight improvements over VGG-16 that has been widely used for face recognition. Since the architecture of ResNet-101 is less complicated than that of Inception-ResNet v1, we only employ ResNet-101 in this paper.

Computational Time. We provide the training and test time of our Multi-Task Multi-Scale ResNet model. In the experiments, we make use of the following computer: 2 Intel Xeon E5-2603 v3 CPUs (1.6GHz and 6 cores for each CPU), 4 Titan X GPUs (12G memory for each GPU), and 128G RAM. When all the 4 GPUs are used parallel, the time of training our model is 71 minutes. Moreover, during test process, the time of processing a pair of test faces is 0.1 second. This means that our Multi-Task Multi-Scale ResNet model can provide very quick diagnosis of both ASD and MDD. This is far more efficient than a psychiatrist who usually makes a diagnosis with at least half an hour.

5 Conclusion

In this paper, we have proposed a novel approach DeepInsight to quick diagnosis of ASD and MDD. To alleviate the scarcity of training data, we have designed a multi-task deep learning model. Moreover, to extract task-aware facial features, we have also induced multi-scale combination into our multi-task deep learning model. The experimental results show that our approach can yield very impressive results in mental disorder diagnosis. We have made the first contribution to learning *both shared and task-aware* deep features for multi-task medical diagnosis, to the best of our knowledge. In the future work, we will collect more facial images to train more robust deep learning models and also extend our DeepInsight approach to other mental disorders. Moreover, we will apply our approach to mental disorder diagnosis with other types of medical data (e.g. brain MRI) other than facial images.

Acknowledgements

The authors would like to thanks the anonymous reviewers and area chairs. This work was partially supported by National Natural Science Foundation of China (61573363 and 61573026), National Basic Research Program of China (2015CB352502), and the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNLQ01). Z. Lu is the corresponding author.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] K. Aldridge, I. D. George, et al. Facial phenotypes in subgroups of prepubertal boys with autism spectrum disorders are correlated with clinical phenotypes. *Molecular Autism*, 2(1):15, 2011.

[3] J. R. Austin, T. N. Takahashi, and Y. Duan. Distinct facial phenotypes in children with autism spectrum disorders and their unaffected siblings. In *International Meeting for Autism Research*, 2012.

[4] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016.

[5] J. Bi, T. Xiong, S. Yu, M. Dundar, and R. B. Rao. An improved multi-task learning approach with applications in medical diagnosis. In *ECML-PKDD*, pages 117–132, 2008.

[6] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, 111 (15):1454–62, 2014.

[7] A. Esteva, B. Kuprel, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[8] S. Fletcher-Watson, V. Leekam, Srbenson, M. Frank, and J. Findlay. Eye-movements reveal attention to social information in autism spectrum disorder. *Neuropsychologia*, 47(1):248–257, 2009.

[9] T. Gehrig and H. K. Ekenel. A common framework for real-time emotion recognition and facial action unit detection. In *CVPR*, pages 1–6, 2011.

[10] Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, pages 6645–6649, 2013.

[11] P. O. Harvey, P. Fossati, J. B. Pochon, R. Levy, G. Lebastard, S. Lehericy, J. F. Allilaire, and B. Dubois. Cognitive control and brain resources in major depression: an fMRI study using the n-back task. *Neuroimage*, 26(3):860–869, 2005.

[12] H. C. Hazlett, H. Gu, et al. Early brain development in infants at high risk for autism spectrum disorder. *Nature*, 542:348–351, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[14] Xiao-Yuan Jing, Hau-San Wong, and David Zhang. Face recognition based on 2d fisherface approach. *Pattern Recognition*, 39(4):707–710, 2006.

[15] A. Krishnan, R. Zhang, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 19:1454–1462, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[17] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[18] J. G. Lee, S. Jun, Y. W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570–584, 2017.

[19] F. Li, L. Tran, K. H. Thung, S. Ji, D. Shen, and J. Li. Robust deep learning for improved classification of AD/MCI patients. In *International Workshop on Machine Learning in Medical Imaging*, pages 240–247, 2014.

[20] J. Liu, Zh. J. Zha, Q. I. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet CNN for person re-identification. In *ACM Multimedia*, pages 192–196, 2016.

[21] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, pages 1749–1756, 2014.

[22] W. Liu, M. Li, and Y. Li. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9 (8):888–898, 2016.

[23] E. Long, H. Lin, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nature Biomedical Engineering*, 1:0024, 2017.

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanad dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010.

[25] F. P. S. Luus, B. P. Salmon, F. Van Den Bergh, and B. T. J. Maharaj. Multiview deep learning for land-use classification. *IEEE Geoscience and Remote Sensing Letters*, 12: 2448–2452, 2015.

[26] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj. Multiview deep learning for land-use classification. *IEEE Geoscience and Remote Sensing Letters*, 12 (12):2448–2452, 2015.

[27] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016.

[28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representation by back-propagation of errors. *Nature*, 323:533–536, 1986.

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] H. I. Suk, S. W. Lee, and D. Shen. Subclass-based multi-task learning for Alzheimer's disease diagnosis. *Frontiers in Aging Neuroscience*, 6:168, 2014.

[32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[33] A. Tripp, H. Oh, J. P. Guilloux, K. Martinowich, D. A. Lewis, and E. Sibille. Brain-derived neurotrophic factor signaling and subgenual anterior cingulate cortex dysfunction in major depressive disorder. *American Journal of Psychiatry*, 169(11):1194–202, 2012.

[34] J. Wang, Q. Wang, et al. Multi-task diagnosis for autism spectrum disorders using multi-modality features: A multi-center study. *Human Brain Mapping*, 38(6):3081–3097, 2017.

[35] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis Decoste, Wei Di, and Yizhou Yu. HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition. In *ICCV*, pages 2740–2748, 2015.

[36] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv Preprint*, abs/1707.00785, 2014.

[37] D. Zeevi, T. Korem, et al. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094, 2015.

[38] F. Zhang, B. Do, and L. Zhang. Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1793–1802, 2016.

[39] Jun Zhang, Yong Yan, and Martin Lades. Face recognition: eigenface, elastic matching, and neural nets. *Proceedings of the IEEE*, 85(9):1423–1435, 1997.

[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, 2016.

[41] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, pages 919–926, 2004.

[42] Z. Zhang, P. Luo, C. L. Chen, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108, 2014.