electronics

MDPI

Article

Deep Multi-Instance Conv-Transformer Frameworks for Landmark-Based Brain MRI Classification

Guannan Li, Zexuan Ji and Quansen Sun *

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; 314106002367@njust.edu.cn (G.L.)
* Correspondence: sunquansen@njust.edu.cn

Abstract: For brain diseases, e.g., autism spectrum disorder (ASD), with unclear biological characteristics, the detection of imaging-based biomarkers is a critical task for diagnosis. Several landmark-based categorization approaches have been developed for the computer-aided diagnosis of brain diseases, such as Alzheimer's disease (AD), utilizing structural magnetic resonance imaging (sMRI). With the automatic detection of the landmarks of brain disease, more detailed brain features were identified for clinical diagnosis. Multi-instance learning is an effective technique for classifying brain diseases based on landmarks. The multiple-instance learning approach relies on the assumption of independent distribution hypotheses and is mostly focused on local information, thus the correlation among different brain regions may be ignored. However, according to previous research on ASD and AD, the abnormal development of different brain regions is highly correlated. Vision Transformers, with self-attention modules to capture the relationship between embedded patches from a whole image, have recently demonstrated superior performances in many computer vision tasks. Nevertheless, the utilization of 3D brain MRIs imposes a substantial computational load, especially while training with Vision Transformer. To address the challenges mentioned above, in this research, we proposed a landmark-based multi-instance Conv-Transformer (LD-MILCT) framework as a solution to the aforementioned issues in brain disease diagnosis. In this network, a two-stage multi-instance learning strategy was proposed to explore both spatial and morphological information between different brain regions; the Vision Transformer utilizes a multi-instance learning head (MIL head) to fully utilize the features that are not involved in the ultimate classification. We assessed our proposed framework using T1-weighted MRI images from both AD and ASD databases. Our method outperformed existing deep learning and landmark-based methods in terms of brain MRI classification tasks.

Keywords: brain disease diagnose; autism spectrum disorder; Alzheimer's disease; vision transformer; multi-instance learning; landmark

check for updates

Citation: Li, G.; Ji, Z.; Sun, Q. Deep Multi-Instance Conv-Transformer Frameworks for Landmark-Based Brain MRI Classification. *Electronics* **2024**, *13*, 980. https://doi.org/10.3390/electronics13050980

Academic Editor: Maria Evelina Fantacci

Received: 12 January 2024
Revised: 25 February 2024
Accepted: 27 February 2024
Published: 4 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Structural Magnetic Resonance Imaging (sMRI) allows clinicians and researchers to noninvasively study the anatomy of the brain in many areas of brain disease diagnosis, e.g., Alzheimer's disease (AD) and autism spectrum disorder (ASD) [1–3]. The classification of brain disease by MRI has been a crucial factor in the development of novel treatments as well as the enhancement of patient outcomes. Furthermore, the diagnosis of ASD is mainly based on behaviors, and so the biomarkers of brain development are rarely known. Therefore, detecting biomarkers from within the brain is of great significance for better intervention. Diagnosis based on MRI has also been an essential contributor to the creation of novel treatments. Many neuroscience studies of children and young adults with ASD have shown patterns of structural and developmental abnormalities in the amygdala [4], hippocampus [4–6], precentral gyrus [7,8], and anterior cingulate gyrus [6]. The detection of the abnormal development of local brain regions may aid in the diagnosis of ASD [7,9]. Thus, utilizing local information details from MRI in anomalous regions could help improve the classification accuracy.

In recent years, the development of deep learning has brought new technical approaches to image-based computer-aided diagnosis, e.g., breast cancer, COVID-19 [10–12]. The 3D Convolutional Neural Network (CNN) with auto-encoder technologies was used to diagnose brain disease in an Alzheimer's dataset [13]. However, features defined at the whole-brain level do not effectively characterize abnormal aspects of brain structure, such as a small size or complex limbic systems in clinical brain regions (i.e., amygdala and hippocampus).

To make more effective use of the local features of brain MRI for brain disease diagnosis, Zhang et al. proposed an image patch-level (intermediate scale between the voxel-level and the region of interest (ROI)) feature detection method for brain disease diagnosis based on brain MRI [14]. Multi-instance learning is an efficient way to exploit the patch-level features of images [12,15,16]. An image is equally divided into N patches, and each image is considered as a bag of patches, with patches considered as instances [12,17]. End-to-end landmark-based deep multi-instance learning (LDMIL) was proposed by Liu et al. for Alzheimer's disease classification/detection [17,18], which converts the classification task into multi-instance learning based on feature landmarks from within the brain, and the full connection layer integrates local structure information to obtain classification results. To address the small datasets issue of early ASD detection, Li et al. proposed a multi-channel convolutional neural network (DE-MC) based on a patch-level data expansion strategy to diagnose infants who are at risk of early ASD. In these MIL-based classification methods, discriminating patches from MR images based on anatomical landmarks were selected, and then the joint learning of the feature representations of the input patches and the subsequent classifier were conducted in a multi-instance convolutional neural network model. However, feature maps in multi-instance learning are captured under the independent identically distributed hypothesis, and the correlation between feature maps is ignored. The independence of different instances may overlook the interconnections between distinct brain regions. However, in the clinical diagnosis of brain diseases, it is critical to observe correlations between abnormal patterns in various regions. For example, there may be a correlation between abnormal cortical thickness and ventricular size in the diagnosis of Alzheimer's disease (AD). Hence, it is crucial to develop a module that links local brain region features together to enhance the accuracy of classification in the MIL-based diagnosis task [19].

The Transformer architecture, introduced by Vaswani et al. [20,21], is now the most prominent model in the domain of natural language processing (NLP). The Transformer model utilizes the attention mechanism, in which tokens attend to each other, enabling the model to capture long-range dependencies more effectively [22,23]. Inspired by the success of the Transformer model in NLP, Dosovitskiy et al. [24] introduced the Vision Transformer (ViT) architecture for image classification applications. Vit has other appealing features; for example, it scales up easily [25] and is robust against corruption [26]. The first stage of the ViT training process is dividing the input image into image patches; each patch is regarded as a word token in the natural language processing task. Unlike the CNN-based network, ViT uses a self-attention mechanism to learn about relationships between patches and performs out-of-stage natural image classification tasks. The self-attention mechanism of ViT effectively captures the relevant features in the image classification while suppressing the disruptive features [19]. However, traditional Transformer sequences are limited by their computational complexity and can only tackle shorter sequences (e.g., less than 1000) [27]. Furthermore, ViT is confronted with the issue of relying on extensive datasets. To enhance the efficiency of ViT on limited data sets, Touvron et al. [28] proposed a data-efficient image Transformer (DeiT). This approach incorporates a convolutional neural network (CNN) to aid in feature extraction. The retrieved features are subsequently fed into the Transformer block. The ViT approach commonly utilizes a solitary class token to deliver the ultimate result for image categorization, whereas the feature tokens derived from patches are often disregarded, Yu et al. [16] proposed an MIL-VT framework for fundus image classification with a multi-instance learning head to make full use of the feature

representations extracted from patches. While ViT-based image classification models excel in various image classification tasks, traditional ViT models are often used for 2D image classification tasks (e.g., ViT [29], TranMIL [19]). However, when it comes to 3D image classification tasks, the computational complexity grows exponentially. However, the majority of brain MRIs are three-dimensional renditions that accurately depict the anatomy of the brain.

To tackle the aforementioned issues, we proposed a landmark-based multi-instance Conv-Transformer (LD-MILCT) framework for brain disease MRI classification. Unlike the conventional ViT model, which divides the input into patches extracted from the entire image, we employ a data-driven approach to extract image patches via landmark extraction. The process of extracting patches that exhibit inter-group variations is capable of capturing local regions that are significant for the classification of brain diseases. A network with two-stage multi-instance learning before/after ViT (LD-MILCT) was proposed for the categorization of brain MRIs. The network employs a multi-instance CNN to capture the local brain structure information provided by patches identified from landmarks. Additionally, a Transformer structure is applied to collect the correlations between local structural and global structural information obtained from all detected landmarks. To optimize the patch instance functionality, the ViT module is equipped with the multi-instance learning head (MIL Head). Compared to the typical full-size patch extraction Transformer-based model, our model is more efficient and successful in capturing global information in brain MRIs. This strategy also reduces computational complexity. Experiments were conducted on datasets of Alzheimer's disease (AD) and Autism Spectrum Disorder (ASD). The experimental results demonstrate that the proposed method exhibits superior performance compared to other ASD classification methods based on MRI.

2. Methods

Our method's schematic diagram is displayed in Figure 1. Anatomical landmarks were identified from MR images using a data-driven approach [14]. Subsequently, patches were retrieved using a patch-level data expansion technique depending on the recognized landmarks. The proposed multi-instance CNN-Transformer framework was used for brain MRI classification.

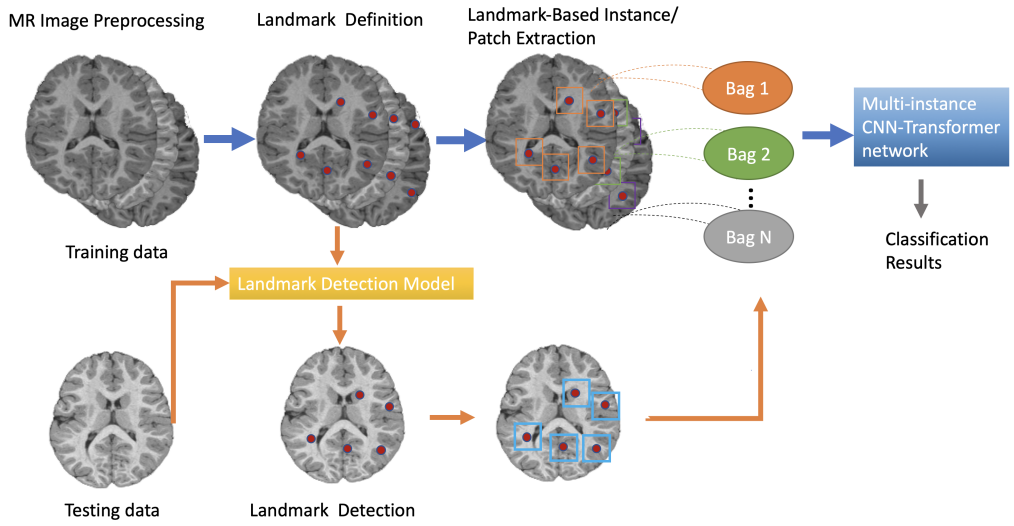


Figure 1. The proposed 3D MRI-based multi-instance Conv-Transformer (LD-MILCT) framework. As for other landmark-based methods, MRI preprocessing and landmarks (shown as the red bullets) detection were necessary; then, landmark-based patches were extracted (an image was donated by a bag of selected feature patches) as inputs for the multi-instance Conv-Transformer network for MRI classification and disease diagnosis.

The preparatory stage is crucial for the ensuing stages of analysis. Skull stripping and histogram matching for T1 MR images were performed using in-house technologies. To adjust for intensity inhomogeneity, each image was resampled to a resolution of $256 \times 256 \times 256$. An artefact often seen in MRIs is that the signal intensity varies smoothly across an image. Variously referred to as RF inhomogeneity, shading artefact, or intensity non-uniformity, it is usually attributed to such factors as poor radio frequency (RF) field uniformity, eddy currents driven by the switching of field gradients, and patient anatomy both inside and outside the field of view. The N3 method (Non-parametric Non-uniformity Normalization) [30] was then applied. To make sure that the dura and skull were removed completely, additional procedures including hand editing and skull stripping were carried out. Ultimately, each skull-stripped image is warped to eliminate the cerebellum using a designated template.

2.1. Data-Driven Anatomical Landmark Identification and Patch Extraction

In clinical diagnosis, brain MRI showed similar overall structure but there were differences in local structures. Previous studies have shown that, in the early stages of AD, there are only small local changes, while the overall structure of the brain changes very little. For ASD diagnosis, the MRI-based biomarker is unclear.

To lighten the load of the MRI-based deep learning classification network as well as capture local features, we first extracted image patches from MRIs using a data-driven landmark detection algorithm [17,31], which identifies and detects brain disease-related landmarks locations [4,17] by looking for statistically significant group differences in training samples. The first step of extracting feature landmarks from the training sample is image registration. We initially chose a T1-weighted image of superior quality to serve as a template. It should be noted that there are many choices of templates, such as Colin27 [32] which refers to the average of 27 registered scans for a single subject. During the preprocessing stage, all training MRIs were aligned with a template by rotation and translation. This ensures that all training pictures are in the same spatial position. The relationship between voxels was established through the process of nonlinear registration. The study utilized the histogram of oriented gradients (HOGs) properties of local brain areas as morphological features to find statistically significant variations between groups [33]. Hoteling's T2 statistic was employed for the statistical comparison. After group comparison, a p-value map was created with each voxel in the template being assigned a significance index p-value. Any voxel in the template with a p-value less than 0.001 is considered a significant landmark.

In practice, a plurality of voxels with p-values less than 0.001 may be concentrated in a small area, so only local minima (whose p-values are also less than 0.001) are defined as landmarks of brain disease in the template.

In the patch extraction stage, for each MRI image, to reduce the influence of registration errors while increasing training samples, a plurality of image patches (as an instances/bag) were randomly extracted within a vicinity around the top L landmarks with the largest group differences, centering on each specific feature landmark. For example, by randomly extracting image patches within a $5 \times 5 \times 5$ cube, 125^L bags/instance can be, hypothetically, generated for each MRI as samples. Each sample is represented by a packet of L image patches as inputs to the sub-CNN in the network structure.

For testing images, we automatically detect landmarks in a new testing image using a shape-constrained regression-forest-based landmark detection method proposed by Zhang et al. [14] using existing landmark information in the training images.

2.2. Multi-Instance Conv-Transformer Neural Networks

The multi-instance Conv-Transformer network was proposed for brain MRI classification as shown in Figure 2a. The input data are L patches taken from L landmarks given a subject. We first ran numerous sub-Conv architectures to reduce the complexity of 3D computations while learning representations of specific patches in each group. The network receives an input MR picture as well as a bag containing L patches, or instances, that were

taken at L landmark points. More precisely, we used a sequence of convolution layers with stride size of 1 and 0-padding to embed L parallel sub-Conv architectures.

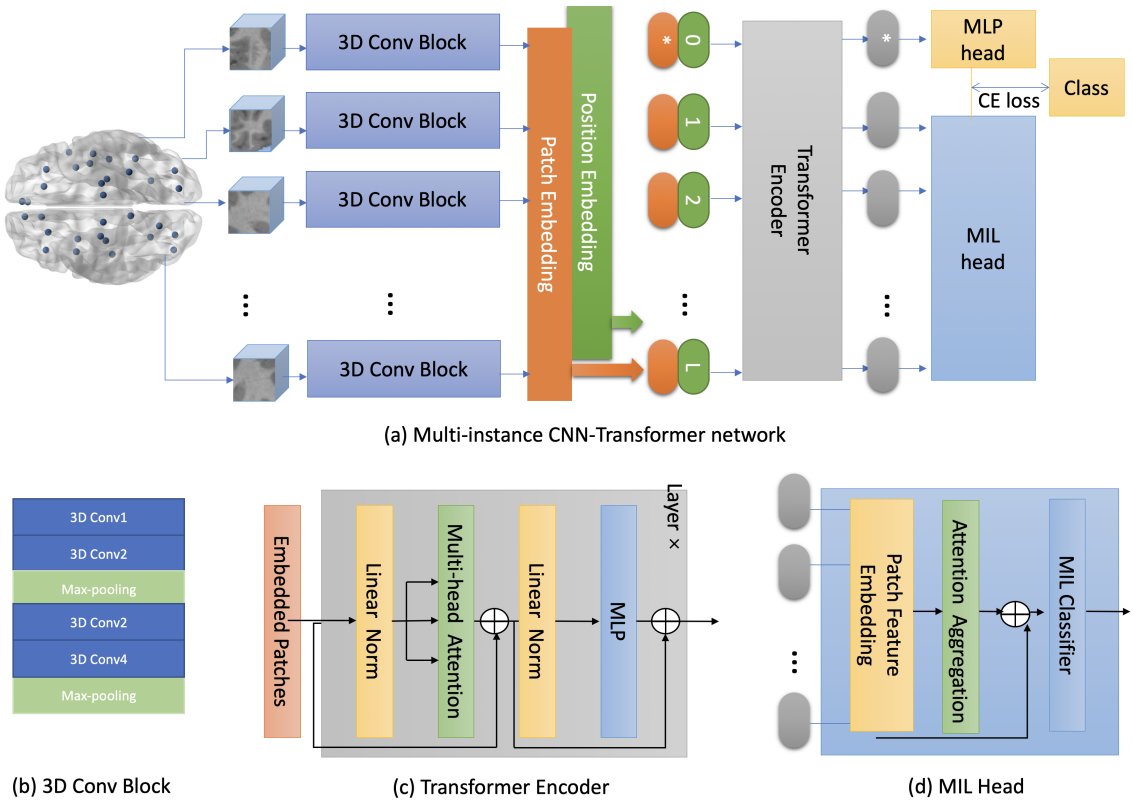


Figure 2. (**a**) Landmark-based multi-instance Conv-Transformer network (LD-MILCT). The blue dots on the brain diagram represent landmarks on which patch extraction is applied. (**b**) The 3D convolution block. The class token is shown as *. (**c**) The Vision Transformer block. (**d**) The Multiple Instance Learning head (MIL) block.

2.2.1. Vision Transformer

The Transformer employs a self-attention method to represent the relationships among tokens in a sequence. Additionally, the inclusion of positional information enhances the utilization of sequential order information. Hence, it is advisable to incorporate the Transformer model into the connected Multiple Instance Learning (MIL) problem [19].

Due to the nuanced nature of structural alterations resulting from brain disease, which can affect several regions of the brain, relying on a single or a few patches is insufficient to accurately depict the overall structural changes occurring in the brain. This differs from typical Multiple Instance Learning, where the picture class can be determined by the estimated label of the most distinguishing patch [17]. The output of the sub-Conv module is transformed into a one-dimensional format and then sent through a linear layer to be injected into D dimensions. The data format is compatible with Transformer networks, as depicted in Figure 2c [24]. The Transformer encoder is composed of alternating layers of multi-headed self-attention (MSA) and MLP blocks [34]. Layer normalization (LN) is implemented before each block, while residual connections are implemented after each block, as depicted in Figure 2c [35].

Location embedding is crucial for encoding spatial location information, as patch embedding alone does not encompass the position of landmarks [24]. As for landmark-based methods, with the nonlinearly aligned MRI processing (with cross-subject voxel-wise correspondence), registration errors may cause interference on the corresponding landmark location of each subject which causes the relative position to not be reliable. Thus, we used standard learnable absolute positional encodings of the landmark position [24]. Previous

studies on ViT show there are no significant performance gains from using more advanced 2D-aware position embeddings [24]. In our method, the patches extracted from landmarks were treated as a sequence. A one-dimensional learnable matrix was employed and we let the network learn the positional information on its own.

The input vector of the standard Transformer is one-dimensional. One possible approach for 3D brain MRI is to extract patches and convert them into one-dimensional representations. The relationship between the bag (as an image) and the instances (as patches) in each MRI can be expressed as $X_n = x_{n,1}, x_{n,2}, \ldots, x_{n,L}$, representing L patches. The dimensions of the MRI patch $x_{n,i}$, extracted using feature landmark, were M × M × M. After feature extraction by the sub-CNN layers, the dimensions were reduced to be smaller, m × m × m, and then the outputs of sub-CNNs were used as inputs for the transformer module. It was then mapped to a D-dimensional vector through the Patch Embedding Layer as shown in Figure 2a. The size of this feature vector is given by $(L + 1) \times D$, where L indicates the number of landmarks chosen from a single MRI (equivalent to the instance number). Both the MLP layer and Muti-attention layer in the Transformer module undergo layer normalization, and there are residual connections between each layer.

2.2.2. Two-Stage Multi-Instance Learning Module

As shown in Figure 2a, the Transformer block has L + 1 input vectors, which include L feature vectors of L patches and an additional class token. The final classification result is determined only by the output of the class token via a multi-layer perceptron (MLP head), disregarding additional patch feature properties. Nevertheless, a single picture block might include significant localized information about the brain area. Hence, it is essential to develop a two-stage multi-instance learning module. This is particularly relevant when considering brain diseases, as their pathology may be distributed in different positions, resulting in varying contributions from different patches. In the MIL scheme, an image is considered as a collection of instances, represented either as pixels or image patches [16,36,37]. The bag–instance relationship depicted in MIL bears a striking resemblance to the image-patch relationship observed in ViT.

The MIL head for the ViT network, as illustrated in Figure 2d and discussed in [16], follows a three-step process. Firstly, it constructs a low-dimensional embedding of ViT output features from a single patch instance. Secondly, it employs an aggregation function to obtain an instance representation. Lastly, it utilizes an image-class (bag-class) classifier to determine the final bag-class probability. The traditional MIL formula is encapsulated and organized within a classification framework, as depicted by the MIL head in Figure 2a.

Patch Feature Embedding. After being processed by the ViT, an instance (image patch) $x_{n,i}$ is transformed into a feature vector $v_{n,i}$ with a dimension of D, represented as $v_{n,i} = T(x_{n,i})$. The feature vectors $v_{n,i}$ are transformed into lower-dimensional embeddings by Linear Layers, Linear Norm, and ReLU activation as shown in Figure 3a.

Attention Aggregation. To enhance the standardization of concealed characteristics and mitigate overfitting issues, in the MIL head, there is an attention module consisting of two linear layers for patch embedding as shown in Figure 3b. The absence of attention aggregation in the model does not exhibit any apparent drawbacks when the input vector is small and the classification problem is straightforward. However, when dealing with a lengthy input feature vector (such as more than 30 local region patches as discussed in this paper), all features are condensed into an intermediate feature vector, resulting in the ineffective transfer of patches' specific features. Consequently, a significant amount of local brain information may be lost. This is another crucial factor that necessitates the introduction of attention aggregation.

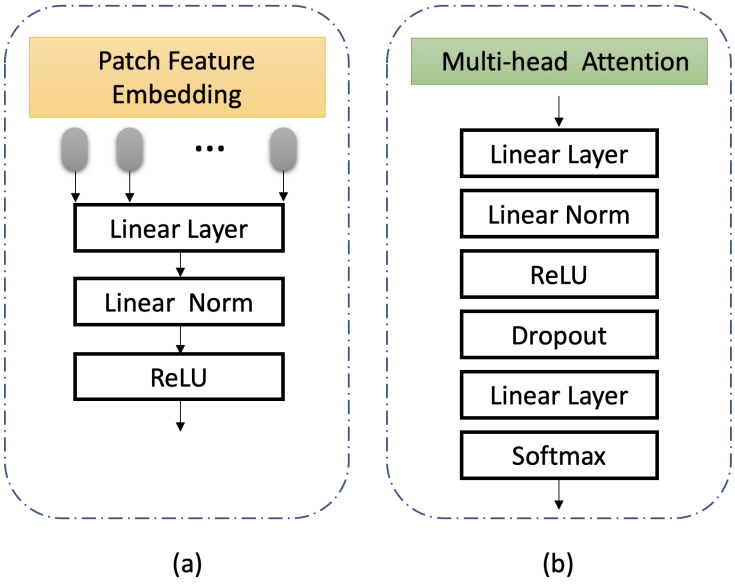


Figure 3. Illustration of the MIL embedding and attention aggregation block. (**a**) Patch feature embedding block. (**b**) Multi-head attention block.

Image-level Classification. The combined patch representation is then inputted into the linear classifier to calculate the ultimate image-level probability.

$$p = w^T R \tag{1}$$

where p represents the prediction probability of the MIL head, $w \in R_{L\times C}$ represents the parameters of the image-level classifier, and C represents the number of groups.

2.2.3. Loss Function

The MLP head and MIL head both utilize the cross entropy loss function to supervise the classification label.

$$Loss(P_1, P_2, W) = -\lambda \sum_{X_n \in \mathbb{X}} log(P_1(y_n|X_n; W)) - (1-\lambda) \sum_{X_n \in \mathbb{X}} log(P_2(y_n|X_n; W)) \tag{2}$$

The variable X_n represents a collection of patches in one MRI. The probabilities $P_1(y_n|X_n; W)$ indicate the likelihood of successfully classifying X_n as class y_n using the class token of ViT or $P_2(y_n|X_n; W)$ for MIL head. These probabilities are determined by the weight parameters W of the network. The variable C represents the number of groups (i.e., ASD/TD) in the MRI classification, while λ is a weight parameter that determines the relative importance of the two classification output heads from ViT and the two-stage multi-instance learning. In practice, λ is typically set at 0.5. During the training phase, the final prediction is obtained by taking a weighted average of the outputs from the two heads, with the weight factor denoted as λ. The network of LD-MILCT is optimized via the Adam optimizer with a batch size of 10 and maximum training epochs of 30, with a learning rate of 10^{-3}.

The Landmark-based multi-instance Conv-Transformer (LD-MILCT) network architecture is an image classification model that operates on patches of 3D MRIs. It is designed to learn both local and global feature representations for 3D MRIs. Specifically, patch-level representations are initially acquired by training numerous sub-Conv structures that correlate to various landmarks in order to capture local structural information situated in distinct regions of the brain. The Visual Transformer layer is used to model the global information represented by various landmarks. The Vision Transformer block enhances the representation of brain structure at the image level by incorporating a multi-instance

learning head (MIL head). Hence, the classifier's learning process can integrate both local and global characteristics of brain MRI.

3. Experiments

3.1. Dataset

Autism Spectrum Disorder datasets The T1-weighted MRIs used for the ASD classification study were obtained from the Autism Brain Imaging Data Exchange (ABIDE), an open-access data repository. These MRIs were collected from 17 international sites without any prior coordination. T1-weighted magnetic resonance images were obtained using 160 sagittal slices. The imaging parameters used were TR/TE = 2400/3.16 ms, and the size was $1 \times 1 \times 1$ mm^3. In this study, we utilized the structural magnetic resonance imaging (MRI) scans of 213 individuals diagnosed with Autism Spectrum Disorder (ASD) and 342 typically developing controls (TD) who were matched in age (ranging from 7 to 49 years, with a median age of 14.4 years across both groups). The experiment included ten-fold cross-validation.

Alzheimer's Disease datasets For the Alzheimer's Disease classification study, we obtained T1-weighted MRIs from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is an open-access datasets. In this study, we utilized structural magnetic resonance imaging (MRI) scans from a total of 308 patients with Alzheimer's disease (AD) and 399 individuals without AD (typical controls or TD subjects). The age range of the participants was between 62 and 85 years, with a median age of 74.9 years across both groups. The experiment included ten-fold cross-validation. To maintain consistency with the classification of Autism Spectrum Disorder (ASD) into two classes, this study did not include the analysis of conversions from mild cognitive impairment (MCI).

3.2. Evaluation

The performance metrics evaluated in this study were accuracy, sensitivity, specificity, and F-score.

Accuracy (*ACC*): Accuracy is the measure of correctly classified 3D MRIs divided by the total number of images in the database. This metric is crucial for assessing the effectiveness and performance of our suggested work. It is characterized by,

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where *TP* represents the True Positives (number of samples correctly identified), *TN* represents True Negatives (samples correctly rejected), *FP* represents false positives (samples incorrectly identified for a given input), and *FN* represents False Negatives (samples that are incorrectly rejected for a given input).

Sensitivity (*SEN*): Sensitivity is a term that can also be referred to as the True Positive Rate (TPR). It represents the ratio of correctly identified positive images to the total number of genuine positive images. The expression for sensitivity is given by the following equation:

$$SEN = \frac{TP}{TP + FN} \tag{4}$$

Specificity (*SPE*): Specificity, also known as the True Negative Rate (TNR), refers to the accurate classification of True Negative samples. Specificity and sensitivity are two contrasting concepts. Specificity is expressed by the following equation:

$$SPE = \frac{TN}{FP + TN} \tag{5}$$

F-score: The *F-score*, also known as the F1 score or F-measure, is an evaluation criterion used to quantify the correctness of a test. The term can instead be described as the calculated

average that takes into account the relative importance of a test's sensitivity and specificity. The *F-score* can be calculated using the following formula:

$$F\text{-}score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{6}$$

3.3. Parameter Analysis

According to the previous studies on landmark-based methods (i.e., LDMIL [17], DE-MC [38]), it is justifiable to select the number of landmarks within the range of 30 to 50. In our experiments, we have chosen to use 40 landmarks. It is important to note that using a larger number of landmarks will increase the number of parameters that need to be optimized in landmark-based methods.

To examine the effects of patch size in the LD-MILCT approach, a group of experiments was performed on the AD datasets by varying the patch size in the set $16 \times 16 \times 16, 20 \times 20 \times 20, 24 \times 24 \times 24, 30 \times 30 \times 30$. By limiting the Vision Transformer block to 1D expansion, we adjusted the convolution kernel size in the sub-Conv block to make the input size of the following Vision Transformer block $10 \times 10 \times 10$ regardless of the input patch size of the sub-Conv. The comparative performance of different patch sizes on AD (Alzheimer's Disease) and TD (Typical Development) classification tasks is presented in Table 1. The most optimal outcomes were achieved using LD-MILCT when utilizing a patch size ranging from 20 to 25. When employing small patches with dimensions of $16 \times 16 \times 16$, the classification performance, as measured by the ACC value of 0.802, is deemed unsatisfactory. This implies that using minute localized areas is insufficient when capturing comprehensive information regarding brain structure. Using big patches in the 3D-Conv block leads to suboptimal outcomes because of the loss of information. Moreover, the utilization of extensive patches would create a substantial computational load.

Table 1. Performance comparison of different patch sizes on AD dataset.

Patch Size	ACC	SEN	SPE	F-Score
$16 \times 16 \times 16$	0.802	0.764	0.851	0.784
$20 \times 20 \times 20$	0.867	0.792	0.786	0.796
$24 \times 24 \times 24$	0.905	0.831	0.940	0.809
$30 \times 30 \times 30$	0.910	0.829	0.947	0.810

3.4. Ablation Studies and Comparison with State-of-the-Art Methods

First, we conducted a comparison between the landmark-based approaches and the CNN-based methods (3D CNN, 3D ResNet [39]). The 3D CNN takes the entire linearly aligned T1-MRI as its input. The network consists of fifteen convolutional layers and three fully connected layers. We have tuned all parameters to provide a fair comparison. The other three landmark-based approaches (LDMIL, and the two variants of the proposed LD-MILCT) use the same sized patch with dimensions of $24 \times 24 \times 24$ and include $L = 40$ landmarks.

The classification performances of the models on the datasets for Alzheimer's Disease (AD) and Autism Spectrum Disorder (ASD) are presented in Table 2. The other three landmark-based methods exhibit a much higher accuracy in both AD and ASD classification when compared to the conventional 3D CNN-based method (ResNet).The reason for this may be that the landmark-based method prioritizes the analysis of local patch-level structural information in the brain.The LD-MILCT algorithm outperformed LDMIL in ASD classification, achieving a significant improvement in accuracy (ACC) of 8.6% The potential cause of this may be that the Vision Transformer block acquires knowledge about the correlation among different brain regions. Given the observable differences in structural abnormalities between individuals with Alzheimer's disease (AD) and typically developing (TD) individuals, just a small number of markers are necessary to differentiate AD participants from TD ones. For the classification of ASD, it is challenging to predict

whether an individual with ASD would transition to typical development (TD) based on just a few indicators, despite the fact that anatomical abnormalities in the ASD brain may be modest and present in several brain regions.

We also conducted a comparative analysis of LD-MILCT and its two variations: (1) LD-MILCT with the MIL head, and (2) LD-MILCT without the MIL head. The experimental results show that the two-stage multi-instance learning module has a 0.3% and 1.5% improvement in classification accuracy on the AD and ASD datasets, respectively, following ViT. The LD-MILCT model demonstrated superior performance in ASD classification, with a notable improvement in accuracy (ACC) of 1.5%. The utilization of an MIL head may enhance the acquisition of relevant local data and enhance the accuracy of categorization. The experimental findings demonstrate that a Transformer block followed by a two-stage multi-instance learning of patch tokens may allow us to efficiently acquire the specifics of nearby brain areas and enhance the categorization outcomes.

Table 2. MRI Classification results with different methods on AD and ASD dataset

Dataset	Model	Combination	ACC	SEN	SPE	F-Score
AD datasets	3D CNN	\	0.680	0.622	0.692	0.664
	ResNet	\	0.830	0.833	0.783	0.726
	LDMIL	\	0.906	0.874	0.930	0.891
	LD-MILCT	\	0.905	0.831	0.940	0.894
	LD-MILCT	with MIL head	0.910	0.903	0.957	0.903
ASD datasets	3D CNN	\	0.562	0.438	0.692	0.611
	ResNet	\	0.575	0.474	0.599	0.211
	LDMIL	\	0.623	0.486	0.601	0.252
	LD-MILCT	\	0.692	0.495	0.686	0.331
	LD-MILCT	with MIL head	0.701	0.508	0.695	0.333

4. Discussion

The proposed LD-MILCT technique shares similarities with, but also exhibits differences from, the multi-instance learning (MIL) methods. The MIL is grounded in the assumption of independent distributions and primarily emphasizes local information, perhaps disregarding the link between multiple brain regions. The LD-MILCT method, unlike the traditional landmark-based brain disease classification method, utilizes the ViT block to learn the relationship between features. This allows it to capture both local and global information from 3D MRIs. As a result, the LD-MILCT method demonstrates improved performance in classifying brain diseases.

The experiment carried out in this paper compared the influence of patch size on the classification effect. The results show that when, the patch size is less than $24 \times 24 \times 24$, the classification result is not ideal. This may be because smaller patch sizes cannot contain enough detailed information. A larger patch size may bring a better classification effect, but it will also bring about an increase in computational complexity. From a comprehensive point of view, It is of great importance to select an appropriate patch size for landmark-based brain disease classification.

Compared to the performance of classifying the AD datasets based on MRI, the proposed framework significantly enhances the accuracy of the classification of the ASD datasets. One possible reason for this is that individuals with Autism Spectrum ASD have more undetectable anomalies in their small brain structure compared to individuals with AD. The utilization of the Transformer module allows the classification model to obtain relevant information about brain regions.

Limitations

There are several limitations to this study. First, in our research, the training sample was limited (i.e., hundreds) for the nature image recognition tasks in which the Vision Transformer demonstrated superior performance on the larger data sets. As far as we

know some physiological traits such as age, gender, height, weight, nutritional status, and education could impact the brain's anatomy and structure [40–42]; this study does not this include additional information or correlations between physiological traits.The image patches were consistently sized across all brain sites, despite the fact that anatomical changes resulting from brain disease can differ across various areas. In the current implementation of our after the 3D-Conv block, the input size of the Vision Transformer block was fixed to a relatively small size, limited by the calculation hardware. This could have had a detrimental impact on the network's generalizing ability.

5. Conclusions

A novel framework for brain 3D MRI classification was proposed in this research, which is based on landmarks and utilizes a multi-instance Conv-Transformer model.

The main contributions of this paper are as follows: Firstly, instead of dividing the entire image into patches, this paper proposes a method of detecting brain disease landmarks in advance. This approach reduces the computational burden of the Transformer module without sacrificing important biomarker information, thereby enhancing the generalization capability of the Transformer model on 3D medical images. Furthermore, this work addresses the inadequacy of multi-instance learning in capturing the link between brain regions by using a Transformer. This allows for the extraction of the feature relationships between brain regions. Thirdly, the implementation of two-stage multi-instance learning after the Transformer module allowed us to capture local information from the patch tokens, which is usually ignored in standard ViT. The proposed technique for classifying 3D brain MRI incorporates the concept that illness signs are interconnected in clinical diagnosis. Our experiments demonstrates that this strategy significantly enhances the accuracy of classifying brain MRI scans.

Author Contributions: G.L. designed the network architecture, analyzed the data, wrote the paper, and experimentally validated the technical routes; Z.J. designed the technical routes, designed the experiments and made revisions to the network; Q.S. made major revisions to the academic content. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China 62372235.

Data Availability Statement: The data source of the ABIDE dataset is the Autism Brain Imaging Data Exchange (ABIDE, http://fcon_1000.projects.nitrc.org/indi/abide); The data source of the ADNI dataset is the Alzheimer's Disease Neuroimaging Initiative (ADNI, https://adni.loni.usc.edu).

Acknowledgments: The data utilized in this publication were acquired from the Autism Brain Imaging Data Exchange (ABIDE) repository and the Alzheimer's Disease Neuroimaging Initiative (ADNI). This manuscript represents the perspectives of the authors and may not necessarily align with the opinions or viewpoints of the individual contributing the original data to the datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Al-Rahayfeh, A.A.; Atiewi, S.; Almiani, M.; Jararweh, M.; Faezipour, M. Utilizing 3D magnetic source imaging with landmark-based features and multi-classification for Alzheimer's Disease diagnosis. *Clust. Comput.* **2023**. [CrossRef]
2. Interagency Autism Coordinating Committee. *2012 Strategic Plan for Autism Spectrum Disorder Research*; Interagency Autism Coordinating Committee: Washington, DC, USA, 2012; p. 104.
3. Kenny, L.; Hattersley, C.; Molins, B.; Buckley, C.; Povey, C.; Pellicano, E. Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism* **2016**, *20*, 442–462. [CrossRef] [PubMed]
4. Li, G.; Chen, M.H.; Li, G.; Wu, D.; Lian, C.; Sun, Q.; Rushmore, R.J.; Wang, L. Volumetric Analysis of Amygdala and Hippocampal Subfields for Infants with Autism. *J. Autism Dev. Disord.* **2022**, *53*, 2475–2489. [CrossRef]
5. Xu, Q.; Zuo, C.; Liao, S.; Long, Y.; Wang, Y. Abnormal development pattern of the amygdala and hippocampus from childhood to adulthood with autism. *J. Clin. Neuroence* **2020**, *78*, 327–332. [CrossRef] [PubMed]

6. Luyster, R.; Gotham, K.; Guthrie, W.; Coffing, M.; Petrak, R.; Pierce, K.; Bishop, S.; Esler, A.; Hus, V.; Oti, R. The Autism Diagnostic Observation Schedule—Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders. *J. Autism Dev. Disord.* **2009**, *39*, 1305–1320. [CrossRef] [PubMed]
7. Wang, L.; Li, G.; Adeli, E.; Liu, M.; Wu, Z.; Meng, Y.; Lin, W.; Shen, D. Anatomy-guided joint tissue segmentation and topological correction for 6-month infant brain MRI with risk of autism. *Hum. Brain Mapp.* **2018**, *39*, 2609–2623. [CrossRef] [PubMed]
8. Schumann, C.M.; Hamstra, J.; Goodlin-Jones, B.L.; Lotspeich, L.J.; Kwon, H.; Buonocore, M.H.; Lammers, C.R.; Reiss, A.L.; Amaral, D.G. The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages. *J. Neurosci.* **2004**, *24*, 6392. [CrossRef]
9. Nacewicz, B.M.; Dalton, K.M.; Johnstone, T.; Long, M.T.; McAuliff, E.M.; Oakes, T.R.; Alexander, A.L.; Davidson, R.J. Amygdala Volume and Nonverbal Social Impairment in Adolescent and Adult Males With Autism. *Arch. Gen. Psychiatry* **2006**, *63*, 1417–1428. [CrossRef]
10. Morais, M.; Calisto, F.M.; Santiago, C.; Aleluia, C.; Nascimento, J.C. Classification of Breast Cancer in Mri with Multimodal Fusion. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Indias, Colombia, 18–21 April 2023; pp. 1–4.
11. Cacciamani, G.E.; Chen, A.B.; Gill, I.S.; Hung, A.J. Artificial intelligence and urology: Ethical considerations for urologists and patients. *Nat. Rev. Urol.* **2024**, *21*, 50–59. [CrossRef]
12. Diogo, P.; Morais, M.; Calisto, F.M.; Santiago, C.; Aleluia, C.; Nascimento, J.C. Weakly-Supervised Diagnosis and Detection of Breast Cancer Using Deep Multiple Instance Learning. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Indias, Colombia, 18–21 April 2023; pp. 1–4.
13. Kruthika, K.R.; Rajeswari; Maheshappa, H.D.; Initiative, A.D.N. CBIR system using Capsule Networks and 3D CNN for Alzheimer's disease diagnosis. *Inform. Med. Unlocked* **2019**, *14*, 59–68. [CrossRef]
14. Zhang, J.; Liu, M.; An, L.; Gao, Y.; Shen, D. Alzheimer's Disease Diagnosis using Landmark-based Features from Longitudinal Structural MR Images. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 1607–1616. [CrossRef]
15. Föllmer, B.; Williams, M.C.; Dey, D.; Arbab-Zadeh, A.; Maurovich-Horvat, P.; Volleberg, R.; Rueckert, D.; Schnabel, J.A.; Newby, D.E.; Dweck, M.R.; et al. Roadmap on the use of artificial intelligence for imaging of vulnerable atherosclerotic plaque in coronary arteries. *Nat. Rev. Cardiol.* **2023**, *21*, 51–64. [CrossRef]
16. Yu, S.; Ma, K.; Bi, Q.; Bian, C.; Ning, M.; He, N.; Li, Y.; Liu, H.; Zheng, Y. MIL-VT: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021.
17. Liu, M.; Lian, C.; Shen, D. *Anatomical-Landmark-Based Deep Learning for Alzheimer's Disease Diagnosis with Structural Magnetic Resonance Imaging*; Springer: Cham, Switzerland, 2019.
18. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 3431–3440.
19. Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; Zhang, Y. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2136–2147.
20. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.M.; Ku, A.; Tran, D. Image Transformer. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 10–14 July 2018.
21. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 12114–12124.
22. Li, J.; Chen, J.; Tang, Y.; Wang, C.; Landman, B.A.; Zhou, S.K. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Med. Image Anal.* **2022**, *85*, 102762. [CrossRef]
23. Lei, T.; Wang, R.; Wan, Y.; Du, X.; Meng, H.; Nandi, A.K. Medical Image Segmentation Using Deep Learning: A Survey. *arXiv* **2020**, arXiv:2009.13120.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
25. Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 11966–11976.
26. Naseer, M.; Ranasinghe, K.; Khan, S.H.; Hayat, M.; Khan, F.S.; Yang, M.H. Intriguing Properties of Vision Transformers. In Proceedings of the Neural Information Processing Systems, Virtual Event, 6–14 December 2021.
27. Xiong, Y.; Zeng, Z.; Chakraborty, R.; Tan, M.; Fung, G.M.; Li, Y.; Singh, V. Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 14138–14148.
28. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; J'egou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual Event, 12–18 June 2020.
29. Dai, D.; Sun, Y.; Dong, L.; Hao, Y.; Sui, Z.; Wei, F. Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers. *arXiv* **2023**, arXiv:2212.10559.
30. Sled, J.G.; Zijdenbos, A.P.; Evans, A.C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* **1998**, *17*, 87–97. [CrossRef]

31. Lyall, A.E.; Shi, F.; Geng, X.; Woolson, S.; Li, G.; Wang, L.; Hamer, R.M.; Shen, D.; Gilmore, J.H. Dynamic Development of Regional Cortical Thickness and Surface Area in Early Childhood. *Cereb. Cortex* **2015**, *25*, 2204. [CrossRef]
32. Holmes, C.J.; Hoge, R.D.; Collins, D.L.; Woods, R.P.; Toga, A.W.; Evans, A.C. Enhancement of MR Images Using Registration for Signal Averaging. *J. Comput. Assist. Tomogr.* **1998**, *22*, 324–333. [CrossRef] [PubMed]
33. Mardia, K. Assessment of multinormality and the robustness of Hotelling's T 2 test. *Appl. Stat.* **1975**, *24*, 163–171. [CrossRef]
34. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
35. Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D.F.; Chao, L.S. Learning Deep Transformer Models for Machine Translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
36. Tang, F.; Ding, J.; Wang, L.; Ning, C.Y.; Zhou, S.K. CMUNeXt: An Efficient Medical Image Segmentation Network based on Large Kernel and Skip Fusion. *arXiv* **2023**, arXiv:2308.01239.
37. Ilse, M.; Tomczak, J.M.; Welling, M. Attention-based Deep Multiple Instance Learning. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
38. Li, G.; Liu, M.; Sun, Q.; Shen, D.; Wang, L. Early Diagnosis of Autism Disease by Multi-channel CNNs. *Mach. Learn. Med. Imaging* **2018**, *11046*, 303–309. [PubMed]
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
40. Littler, N. Brain development education pathways. *Br. J. Sch. Nurs.* **2016**, *11*, 98–99. [CrossRef]
41. Cheatham, C.L. Nutritional Factors in Fetal and Infant Brain Development. *Ann. Nutr. Metab.* **2020**, *75*, 20–32. [CrossRef]
42. Giedd, J.N.; Blumenthal, J.D.; Jeffries, N.O.; Castellanos, F.X.; Liu, H.; Zijdenbos, A.; Paus, T.; Evans, A.C.; Rapoport, J.L. Brain development during childhood and adolescence: A longitudinal MRI study. *Nat. Neurosci.* **1999**, *2*, 861–863. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.