WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 4, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n4.09
Journal homepage: https://wepub.org/index.php/IJCSIT/index

ESC-Net: A Lightweight EfficientNetV2-Based Framework with Coordinate Attention and Global Context Enhancement for Alzheimer's Disease Auxiliary Diagnosis

Xinru Xue, Yan Gao, Zihao Yang, Pan Zhu, Hangfan Zhou, Jiahao Piao, Min Liu *

College of Information Engineering, Henan University of Science and Technology, Luoyang, China
*Corresponding Author: Min Liu

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder, and early identification is crucial for delaying disease progression. Existing MRI-based diagnostic methods still face challenges in computational efficiency, model lightweighting, and multi-stage classification accuracy. This paper proposes a lightweight auxiliary diagnostic model named ESC-Net based on an improved EfficientNetV2 for three-stage classification of cognitively normal (CN), mild cognitive impairment (MCI), and AD. The model adopts a stage-wise heterogeneous convolutional design. Shallow layers use FusedMBConv to enhance training efficiency, while deep layers integrate a Coordinate Attention (CA) mechanism into MBConv to capture both channel relationships and spatial dependencies, improving localization of key pathological regions such as the hippocampus. A progressive stochastic depth regularization strategy is introduced to mitigate overfitting in small-sample medical imaging data. Experimental results on the ADNI dataset show that the proposed model achieves a sensitivity of 99.54% for AD diagnosis, an accuracy of 98.67% for early MCI identification, and a specificity of 98.02% for distinguishing MCI from CN. Compared with traditional deep convolutional networks, this model significantly reduces computational complexity while maintaining excellent classification performance, demonstrating promising potential for clinical application and mobile deployment.

KEYWORDS

Alzheimer's disease; EfficientNetV2; Coordinate attention; Lightweight model; MRI classification

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder with an insidious onset, clinically characterized by memory impairment, aphasia, apraxia, agnosia, and executive dysfunction [1]. With the increasing trend of global population aging, the prevalence of AD has shown explosive growth, imposing a heavy burden on public health systems [2]. Studies have shown that early detection and intervention can effectively inhibit the progression from mild to severe stages of the disease [3]. Currently, magnetic resonance imaging (MRI), as a non-invasive examination technique, provides the possibility for early diagnosis by identifying structural abnormalities of brain soft tissues (e.g., hippocampal atrophy, cortical thinning, etc.) [4]. However, manual reading by radiologists has limitations such as strong subjectivity and low efficiency.

In recent years, deep learning techniques have been widely used in computer-aided diagnosis of medical images and have achieved good results [5]. Researchers have made substantial efforts to develop neuroimaging techniques and auxiliary diagnostic strategies. Liu et al. [6] proposed a CNN-based multi-modal deep learning framework for joint automatic hippocampus segmentation

and AD classification, achieving an accuracy of 72.2%. However, the overall classification accuracy of this method is relatively low, making it difficult to meet the high-reliability requirements of clinical diagnosis. Moreover, the joint task design increases the complexity of model training, which may lead to interference between tasks. Khatri and Kwon [7] proposed a lightweight convolutional-attention hybrid model that incorporates inverted residual units and a lightweight multi-head self-attention mechanism, achieving a multi-class classification accuracy as high as 94.31% on the ADNI dataset. However, its performance heavily relies on specific data distributions, and its generalization ability across different centers and scanners has not been fully validated, posing a risk of overfitting in more complex clinical scenarios. Plant et al. [8] used data mining algorithms combining multiple classifiers such as support vector machine (SVM), Bayesian statistics, and voting feature intervals (VFI) for AD analysis. Although this approach has strong interpretability, feature extraction depends on manual design, making it difficult to automatically learn deep pathological features from images, and the ensemble of multiple classifiers makes the overall model relatively complex.

Although the above CNN-based methods have improved diagnostic performance to some extent, their feature extraction and attention modeling still have shortcomings, making it difficult to fully capture pathological features that combine channel importance and precise spatial location in MRI images. Moreover, mainstream visual attention mechanisms (e.g., Squeeze-and-Excitation) typically focus on modeling inter-channel relationships but fail to explicitly capture spatial location information [11]. In AD MRI images, key pathological features (such as hippocampal atrophy) have clear and fixed anatomical locations [12]. Therefore, an attention mechanism that can simultaneously model channel importance and long-range spatial dependencies is expected to more precisely guide the model to focus on relevant regions, thereby improving the ability to identify subtle lesions. The Coordinate Attention (CA) mechanism proposed by Hou et al. [14] is a representative example of such an approach, embedding position information into channel attention so that the network can focus not only on "which features are meaningful" but also on "where the features are located."

Based on this, this paper proposes a lightweight auxiliary diagnostic system using EfficientNetV2 as the backbone network, integrating stage-wise heterogeneous design, a coordinate attention mechanism, and a progressive stochastic depth strategy, aiming to achieve accurate and efficient identification of Alzheimer's disease (AD). The design and advantages of the proposed model are mainly reflected in the following three aspects:

(1) To address the difficulty of a single traditional convolutional structure in balancing local texture and global semantic modeling, this paper adopts the stage-wise heterogeneous design of EfficientNetV2. Specifically, computationally more efficient FusedMBConv is used in shallow layers to extract local texture features, while MBConv is used in deep layers to model high-level semantic information. This design improves the hierarchy and adaptability of feature representation while keeping the model lightweight, which is more consistent with the logic of medical image analysis from coarse to fine.

(2) To address the problem that mainstream attention mechanisms (e.g., SE) focus only on the channel dimension and ignore spatial location information, this paper introduces the Coordinate Attention (CA) mechanism. CA embeds position information into channel attention weights by performing one-dimensional pooling and encoding along the height and width directions, and integrates CA into the residual connection branch of the MBConv module. This mechanism endows the model with both channel-wise selective attention and precise spatial localization capabilities, effectively enhancing sensitivity to pathological changes in key brain regions such as the hippocampus.

(3) To overcome the limited receptive field and insufficient generalization ability of lightweight models, this paper inserts a Global Context Enhancement (GCE) module before the head feature projection. Through multi-scale adaptive pooling and attention fusion, the GCE module effectively models cross-scale pathological features ranging from local subtle lesions to widespread brain

atrophy. At the same time, a progressive stochastic depth strategy is introduced to dynamically adjust the dropout probability of each module, enhancing generalization performance while maintaining the advantages of model lightweighting, and mitigating the risk of overfitting caused by limited medical image samples.

2. RELATED THEORIES AND BASIC ARCHITECTURE

The core architecture of the Alzheimer's disease auxiliary diagnostic model proposed in this study evolves from the lightweight EfficientNetV2 network [15]. This architecture achieves deep integration and optimization of convolution operators, residual connections, and coordinate attention mechanisms in the field of deep learning.

2.1. Inverted Residual Structure and MBConv Operator

The introduction of the residual structure (ResNet) provides an effective solution to the vanishing gradient problem in deep networks. He et al. [3] introduced identity mapping, which significantly increased the achievable depth of networks and laid the foundation for the design of deep convolutional neural networks. Subsequently, to address the limited computational resources of mobile devices, Sandler et al. [16] made important improvements to the traditional residual module in MobileNetV2 and proposed an "inverted residual" design — adopting a reverse strategy of "expand first, then reduce". This expands the channel dimension before feature extraction to enhance feature representation capability, and then compresses it back to the original dimension. This design effectively reduces parameter redundancy and computational cost while maintaining feature extraction capability. On this basis, Tan et al. [1] further integrated MBConv into the EfficientNet series of networks and introduced neural architecture search to achieve coordinated scaling of channel number, depth, and resolution, striking a balance between accuracy and efficiency in image classification tasks.

Inspired by the above studies, this paper introduces MBConv as the core feature extraction module to meet the need for efficient analysis of brain MRI images on lightweight devices. Compared with standard convolutional modules, MBConv combines depthwise separable convolution and channel attention mechanisms, enabling the extraction of richer non-linear features at very low computational cost, which is particularly suitable for brain MRI images with complex structures and subtle textures. This design provides a feasible technical foundation for achieving high-accuracy, low-latency brain image classification on clinical lightweight devices.

2.2. Identity Mapping

Deep convolutional networks face the problems of vanishing gradients and network degradation when increasing depth, making the model difficult to train effectively. To address this issue, He et al. [16] introduced an identity mapping path. The core idea is to directly pass the input to the output and add it to the features after nonlinear transformation. Its mathematical form can be expressed as:

$$y = x + F(x) \tag{1}$$

This allows the network to only fit the residual between the input and the output when learning the target function, rather than the complete mapping, thereby reducing the optimization difficulty. More critically, during the backpropagation of gradients, the gradient can be directly propagated back to the shallow layers through the identity path, effectively avoiding the vanishing gradient problem caused by increased network depth.

As medical image analysis tasks become more advanced, models not only require deeper structures but also the ability to capture subtle pathological features. In the diagnosis of Alzheimer's disease (AD), fine anatomical changes such as sulcal texture and hippocampal structure are crucial for diagnosis. However, simply stacking depth and using standard residual structures may still lead to the dilution of low-level detail information during transmission due to multiple nonlinear transformations.

To address this, the proposed architecture further optimizes the information transmission mechanism while inheriting the idea of residual connections. By retaining the identity mapping alongside the convolutional path, the model learns residual information rather than the original mapping, ensuring that underlying anatomical details are losslessly transmitted to high-level semantic layers. This enhances the model's sensitivity to pathological features while maintaining training stability.

3. METHOD

3.1. Model Overall Architecture

This paper proposes a lightweight classification framework named ESC-Net based on the EfficientNetV2 backbone network. The overall architecture is shown in Figure 1. ESC-Net adopts a stage-wise serial architecture to enhance the model's ability to represent multi-scale structural information and global dependencies with low computational overhead. Specifically, the model is constructed in a serial manner: shallow layers introduce the FusedMBConv module to prioritize capturing spatial details and improve the efficiency of early feature extraction; deep layers adopt the LightMBConvWithAttention module embedded with a coordinate attention mechanism, which enhances spatial localization of key lesion areas such as the hippocampus through direction-aware attention weights; before the head feature projection, a global context enhancement module is inserted to aggregate whole-brain feature interactions via multi-scale pooling and adaptive fusion, compensating for information loss during downsampling. At the same time, the model introduces a progressive stochastic depth regularization strategy, linearly increasing the path dropout probability with network depth, effectively mitigating the overfitting problem in small-sample medical imaging. The overall framework is trained in an end-to-end manner, balancing the requirements of lightweight deployment and high-accuracy recognition.

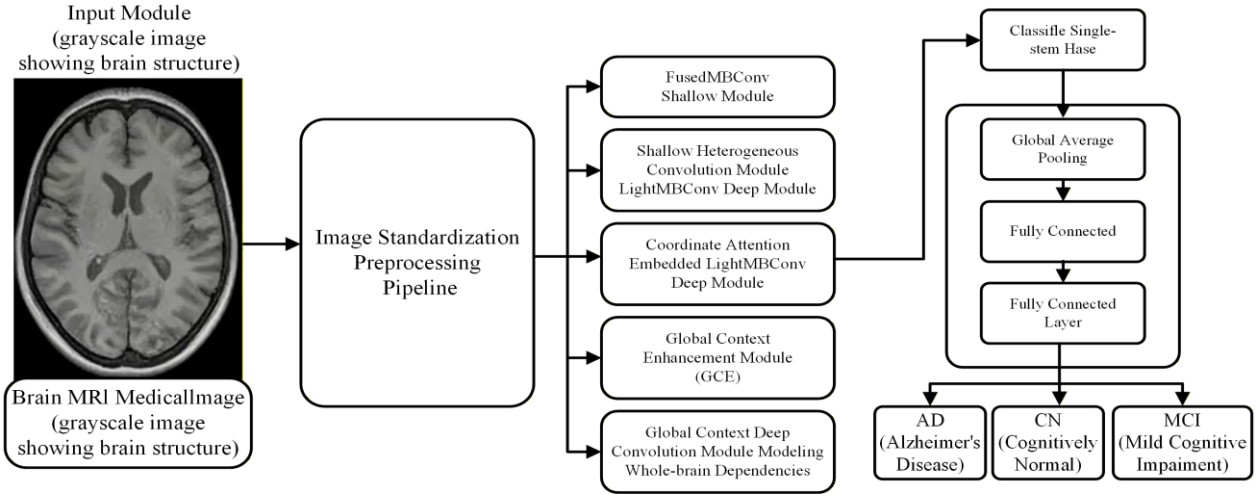

Figure 1. Model Architecture Diagram

3.2. Compound Scaling Strategy of EfficientNetV2

Traditional network scaling methods only expand a single dimension—depth, width, or resolution—which easily leads to rapid saturation of accuracy gains. Moreover, the three dimensions are interdependent: high resolution requires a deeper network to increase the receptive field and also

wider channels to capture fine-grained features. To address this issue, ESC-Net inherits and improves the idea of non-uniform compound scaling. Its structural diagram is shown in Figure 2. Specifically, the model tailors the scaling allocation according to the differences in feature scales presented by medical images at different network depths. In the shallow stages, the model focuses on adjusting the width and adopts the more computationally efficient FusedMBConv module to improve hardware parallelism and early feature extraction speed, giving priority to capturing local texture and edge details of key regions such as the hippocampus. In the middle stages, it balances width and depth, taking into account both feature representation and computational cost, enabling the model to gradually abstract pathological features from local to global scales. In the deep stages, it focuses on increasing depth and embeds a coordinate attention mechanism to enhance spatial perception of subtle pathological changes such as brain atrophy, achieving high-level semantic modeling of the whole-brain structure. In addition, a progressive learning strategy is adopted, gradually increasing the input resolution during training and dynamically adjusting the regularization strength to avoid the risk of overfitting on small samples caused by high-resolution images. Through this differentiated multi-dimensional collaborative scaling, ESC-Net effectively solves the problems of lightweight models in AD diagnosis—namely, the difficulty of balancing local fine features with global macro features, and insufficient capture of multi-scale pathological information—while controlling the number of parameters and computational cost, achieving a balance between efficiency and accuracy.

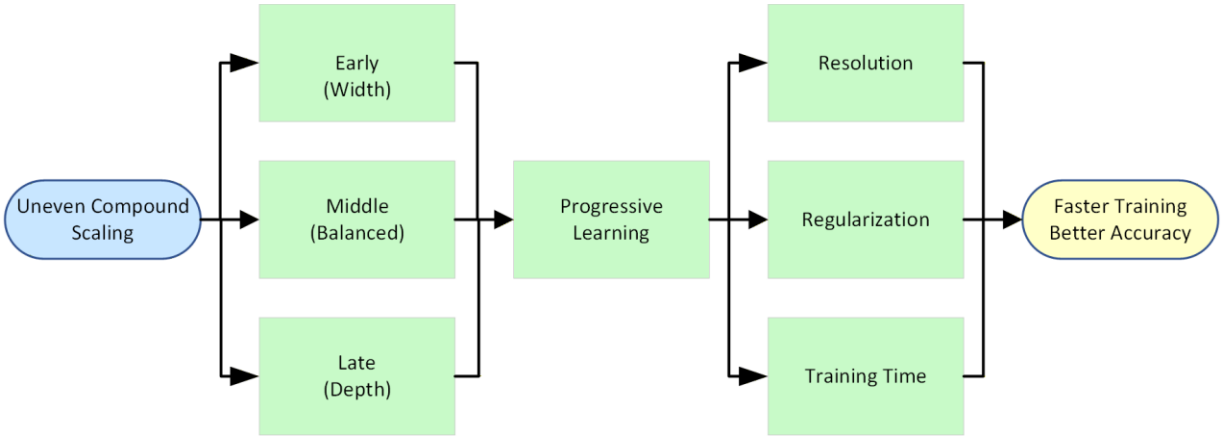

Figure 2. Compound Scaling Strategy Structure Diagram

The formula is as follows:

$$\begin{aligned} depth: d &= \mathrm{d}_0 \cdot \alpha^{\phi} \\ width: w &= w_0 \cdot \beta^{\phi} \\ resolution: r &= r_0 \cdot \gamma^{\phi} \end{aligned} \tag{2}$$

$$\alpha \bullet \beta^2 \bullet \gamma^2 \approx 2 \tag{3}$$

Where d_0 , w_0 , r_0 represent the depth, width, and input resolution of the baseline network, respectively;ϕis the compound scaling coefficient that controls the overall computational resource budget; α,β,γare the scaling factors to be searched, determining the growth rates of depth, width, and resolution, respectively; the constraint condition $\alpha \bullet \beta^2 \bullet \gamma^2 \approx 2$ ensures that when ϕincreases by 1, the total computational cost (FLOPs) of the model approximately doubles.

3.3. Introduction and Balance of Heterogeneous Convolution Modules

In conventional convolutional neural networks, a single type of convolution operator is often used throughout the entire network, making it difficult to simultaneously meet the demand for computational efficiency in shallow layers and the demand for spatial perception in deep layers. As a result, lightweight models struggle to balance local detail preservation and global semantic modeling in AD diagnosis. To address this issue, this paper proposes a stage-wise heterogeneous convolution module design strategy, with the structural diagram shown in Figure 3. This strategy adopts differentiated designs according to the different requirements of shallow and deep layers. In the shallow layers, the more computationally efficient FusedMBConv is introduced, which merges depthwise convolution and pointwise convolution into a conventional convolution, reducing memory access and improving hardware parallelism. This preserves the primary anatomical structures of key regions such as the hippocampus while ensuring feature extraction speed. In the deep layers, MBConv embedded with a coordinate attention mechanism is adopted. Through the process of dimension expansion, depthwise convolution, attention weighting, and then dimension reduction, it enhances the spatial localization ability for subtle pathological changes such as brain atrophy while maintaining lightweight characteristics. Through this stage-wise heterogeneous design, ESC-Net effectively solves the problem that lightweight models cannot easily balance early local feature extraction and deep global semantic modeling in AD diagnosis, achieving a balance between feature extraction efficiency and representational capability while controlling the number of parameters and computational cost.

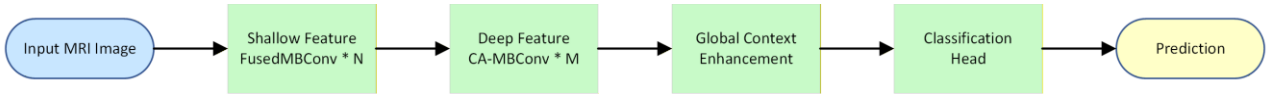

Figure 3. Structure Diagram of the Heterogeneous Convolution Module

The formula is as follows:

$$\mathrm{F}_{FMB}(X) = \mathrm{Pr}oj(Act(BN(Conv_{k\times k}(X)))) + 1_{sc} \cdot X \tag{4}$$

$$F_{MB+CA}(X) = \mathrm{Pr}oj(CA(DWConv(Expand(X)))) + 1_{sc} \cdot X \tag{5}$$

Where X is the input feature map, represents the standard convolution operation, BN denotes batch normalization, Act denotes the SiLU activation function, Proj denotes the 1×1 projection convolution used for channel adjustment, 1sc is the residual connection indicator, $Expand(X) = Act(BN(Conv_{1\times1}(X)))$ is the expansion operation, DWConv denotes depthwise separable convolution, and CA denotes the coordinate attention mechanism.

3.4. Coordinate Attention Mechanism

The traditional Squeeze-and-Excitation (SE) attention mechanism compresses spatial information through global pooling, which fails to capture positional information and makes it difficult to precisely locate regions such as the hippocampus that have clear anatomical locations. To address this issue, this paper embeds the coordinate attention mechanism into the residual connection branch of the MBConv module, as shown in Figure 4. Specifically, in the deep MBConv module, by placing the coordinate attention between the depthwise convolution and the projection convolution, the model performs one-dimensional pooling and encoding along the height and width directions, respectively, in the expanded high-dimensional feature space, generating direction-aware attention weights, which are then multiplied element-wise with the original features. This design enables the network to simultaneously capture inter-channel dependencies and precise spatial positional relationships, enhancing the spatial localization capability for key brain regions such as the hippocampus without

significantly increasing computational cost. It effectively alleviates the problem of losing subtle pathological features in lightweight models caused by limited receptive fields, thereby improving the sensitivity and accuracy of early AD diagnosis.

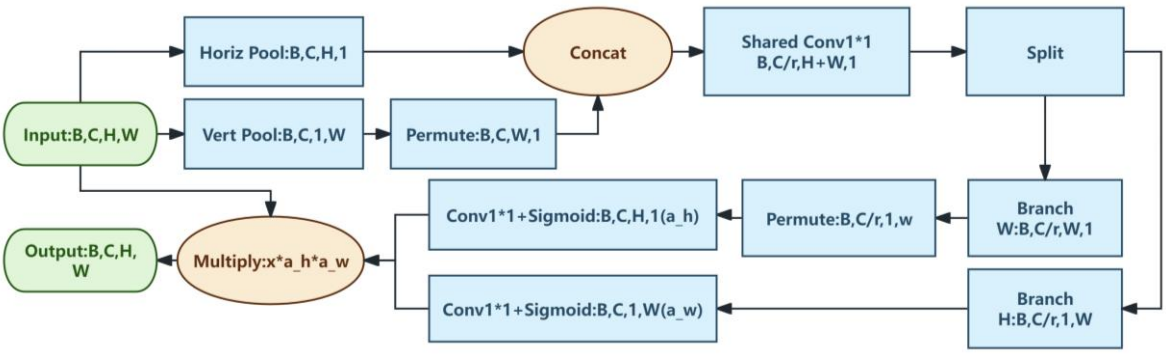

Figure 4. Structure Diagram of the Coordinate Attention Mechanism

The formula is as follows:

$$
\begin{aligned}
&z_h = \frac{1}{W}\sum_{i=1}^{W} x_{:,i,:}, z_w = \frac{1}{H}\sum_{j=1}^{H} x_{:,j,:} \\
&f = \delta(F_1([z_h, z_w])) \\
&g_h = \sigma(F_h(f_h)), g_w = \sigma(F_w(f_w)) \\
&y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j)
\end{aligned}
\tag{6}
$$

$x_c(i,j)$ is the value of the input feature map at channel c and position (i, j), z_c^h and z_c^w are the feature vectors pooled along the height and width directions, respectively, F_1, F_h, and F_w are 1×1 convolution operations, δ is the Hardswish activation function, σ is the Sigmoid activation function, and g^h and g^w are the generated attention weights along the height and width directions.

3.5. Global Context Enhancement Module

This paper proposes a Global Context Enhancement (GCE) module [11] that enhances the model's ability to understand global structure while maintaining lightweight advantages through multi-scale pooling and adaptive fusion strategies. Its structure is shown in Figure 5. This method first uses adaptive average pooling at multiple scales to extract global context features at different granularities. After upsampling to restore the original resolution, lightweight convolutions are applied to extract context representations at each scale. Subsequently, an adaptive attention mechanism is used to perform weighted fusion of the multi-scale contexts. Finally, the enhanced features are combined with the original features via a residual connection. This design enables the model to simultaneously capture multi-scale pathological features ranging from local to global, and the module is lightweight and flexible, allowing seamless integration into existing convolutional networks. In the proposed model, the GCE module is inserted before the head feature projection of EfficientNetV2, complementing the coordinate attention mechanism—coordinate attention focuses on position-sensitive spatial-channel relationships, while GCE focuses on global context modeling. Together, they synergistically enhance the model's multi-scale perception of AD pathological features

Figure 5. Structure Diagram of the Global Context Enhancement Module

The formula is as follows:

$$
\begin{aligned}
&C_k = f_{\theta_k}(Up(Pool_k(x))), k \in \{1,2,4\} \\
&C_{ori} = f_{\phi}(x) \\
&W = Soft\max(F_{fusion}(C_{ori} \oplus C_1 \oplus C_2 \oplus C_4)) \\
&C_{fused} = \sum_k W_k \cdot C_k \\
&A = \sigma(F_{enchance}(x \oplus C_{fused})) \\
&y = x + \alpha \cdot (x \bullet A)
\end{aligned}
\tag{7}
$$

Where x is the input feature map; $Pool_k$ denotes adaptive average pooling with kernel size k × k (k = 1, 2, 4); Up denotes the upsampling operation; f_{θ_k} and f_{ϕ} are context feature extractors composed of 1×1 convolutions; C_k and C_{ori} represent the contextual features after multi-scale pooling and at the original scale, respectively; $\oplus$ denotes concatenation along the channel dimension; F_{fusion} is a convolution module that generates fusion weights; W represents the adaptive fusion weights for multi-scale contexts; C_{fused} is the multi-scale context after weighted fusion; $F_{enchance}$ is a convolution module that generates enhancement weights;σ is the Sigmoid activation function; A is the feature enhancement weight; α is a learnable scaling factor; $\bullet$ denotes element-wise multiplication; and y is the final output.

4. EXPERIMENTAL ENVIRONMENT AND DATA PREPROCESSING

4.1. Dataset Acquisition and Sample Distribution

The experimental data in this study are sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) public database. This database is a benchmark for global neuroimaging research and provides rich structural magnetic resonance imaging (sMRI) data. To validate the discriminative ability of the improved model for different stages of the disease, a three-class dataset was constructed, covering:

CN (Cognitive Normal): Healthy subjects with normal cognition. MCI (Mild Cognitive Impairment): Subjects showing early memory impairment but not meeting the criteria for dementia. AD (Alzheimer's Disease): Patients with confirmed diagnosis and obvious brain atrophy characteristics.

To eliminate non-pathological interference introduced by uneven illumination, equipment differences, and complex grayscale levels in the original MRI images, this study constructed an automated preprocessing pipeline including modality calibration (RGB three-channel conversion), scale unification (bilinear interpolation normalization), and pixel standardization (mean-variance normalization). This enhances the redundant representational capacity of features, adapts to pre-trained weights, and ensures gradient stability during the initial training stage.

4.2. Training Environment Parameter Configuration and Evaluation Criteria

The model is trained using the Adam optimizer with an initial learning rate of 0.001 and a dynamic decay strategy, and a batch size of 32. Additionally, a dual regularization mechanism consisting of global dropout (p = 0.2) and progressive stochastic depth (drop_connect_rate = 0.2) is introduced to mitigate overfitting in small-sample medical imaging data and to enhance the model's generalization ability.

4.3. Evaluation Metrics

We use four metrics to evaluate the model's performance, including Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-Score (F1). Among these metrics, Accuracy is determined by the proportion of true positives (TP) and true negatives (TN) in the total samples; Precision is determined by the proportion of true positives among all positive results; Recall is determined by the proportion of true positives in the sum of true positives and false negatives (FN); and the F1-Score is determined by the proportion of twice the true positives in the sum of twice the true positives and the total number of incorrect classifications. The mathematical expressions for the above metrics are as follows:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ \Pr ecision &= \frac{TP}{TP+FP} \\ \mathrm{Re}call &= \frac{TP}{TP+FN} \\ F1-Score &= \frac{2TP}{2TP+FP+FN} \end{aligned} \tag{8}$$

Among them, TP represents true positives, i.e., the number of positive samples correctly identified. FN represents false negatives, i.e., the number of negative samples incorrectly identified. FP represents false positives, i.e., the number of positive samples incorrectly identified. TN represents true negatives, i.e., the number of negative samples correctly identified.

5. ANALYSIS AND DISCUSSION OF EXPERIMENTAL RESULTS

5.1. Training Dynamics and Convergence Evaluation

As shown in Table 1, the proposed model achieves excellent classification performance across all three categories. For the AD category, the precision, recall, and F1-score all reach 1.0000; the F1-score for the CN category is 0.9818, and for the MCI category it is 0.9878, with an overall accuracy of 98.87%. These results indicate that the model achieves balanced performance across different categories without significant bias. This outcome further validates the effectiveness and robustness of the proposed method in AD diagnosis tasks, providing reliable support for clinical auxiliary diagnosis.

Table 1. Classification Results of the Test Set

	precision	recall	F1-score	support
AD	1.0000	1.0000	1.0000	437
CN	0.9787	0.9852	0.9818	607
MCI	0.9900	0.9856	0.9878	903
Macro avg	0.9896	0.9903	0.9899	1947

Figure 6 shows the accuracy and loss function curves of the model. The accuracy continuously increases and stabilizes, while the loss continuously decreases and converges smoothly, indicating that the model has good generalization performance during training.

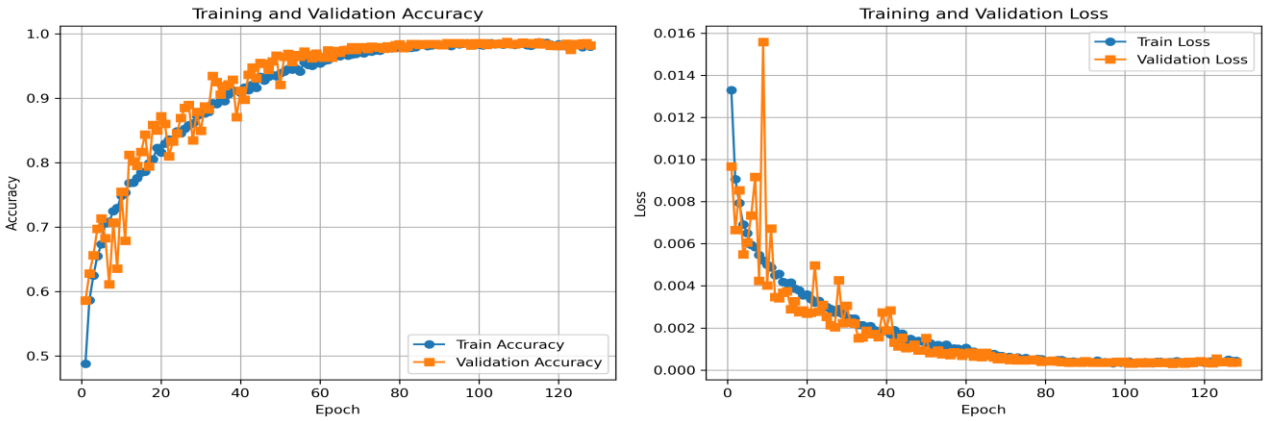

Figure 6. Accuracy and Loss Curves

To further verify the reliability of the model in clinical auxiliary diagnosis, this paper conducts a detailed quantitative evaluation of the three-class classification results using a confusion matrix. Figure 7 shows the confusion matrix and ROC curve, indicating that the model achieves extremely high classification accuracy for the three categories (AD, CN, MCI), with excellent classification performance and strong diagnostic capability.

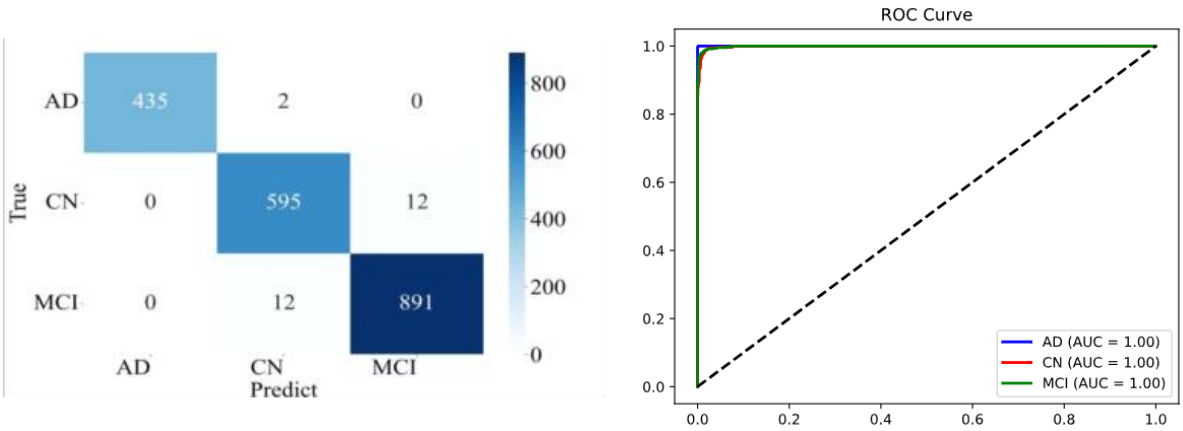

Figure 7. (a) Confusion Matrix (b) ROC Curve

5.2. Ablation Experiment

Ablation experiments were conducted on the proposed model, and the results are shown in Table 2. Here, Baseline denotes the original structure without any improvements. Model 1 denotes the model without the mixup algorithm, using traditional cross-entropy (CE) as the loss function. Model 2 denotes the model without the coordinate attention (CA) mechanism but with the global context enhancement (GCE) module. Model 3 denotes the model with the CA mechanism but without the GCE module. Model 4 denotes the model with both the CA mechanism and the GCE module. The comparison between the proposed model and other models is shown in Table 3.

According to Tables 1 and 2, on the dataset used in this paper, the model employing the proposed method achieves 98.92% accuracy, 99.54% sensitivity, and 98.77% F1-score, with an AUC value of 99.95%. Compared with the traditional networks ResNet34 and EfficientNetV2, the number of parameters is reduced by approximately 5.5% and 4.0%, respectively, while significantly outperforming them in accuracy, AUC, sensitivity, F1-score, and other metrics. This indicates that the proposed method has better feature extraction and classification capabilities for this dataset. Compared with the lightweight models AlexNet and Vision Transformer, the proposed model achieves substantial improvements across all metrics while having far fewer parameters, demonstrating a good balance between lightweight design and high accuracy. Comprehensive

comparison shows that the proposed method achieves optimal classification performance while maintaining lightweight characteristics, validating the effectiveness of the coordinate attention and global context enhancement modules in feature representation and classification capability.

Table 2. Ablation study of the proposed model

Method	ACC (%)	SEN (%)	F1 Score (%)	AUC (%)
Baseline	96.02	96.09	95.43	95.75
Model 1	98.66	98.66	98.67	98.69
Model 2	98.56	98.77	98.61	99.90
Model 3	98.87	99.03	98.99	99.93
Model 4	98.92	99.54	98.77	99.95

Table 3. Comparison Results Between the Proposed Model and Other Models

Models	ACC/%	SEN/%	F1 Score/%	AUC/%	Parameters	FLOPs
ResNet34	94.10	94.46	94.00	94.22	21.79	3.67
AlexNet	86.34	86.19	86.35	86.26	61	0.71
EfficientNet	93.48	94.40	93.48	93.93	57.3	0.41
EfficientNetV2	96.02	96.09	95.43	95.75	21.45	2.90
Vision Transfoemer	93.50	93.11	92.24	92.67	53.3	11.28
ESC-Net	98.92	99.54	98.77	99.95	20.63	2.91

6. CONCLUSION

This paper proposes a lightweight auxiliary diagnostic scheme based on an improved EfficientNetV2 to address the challenges of weak pathological features and small-sample overfitting in Alzheimer's disease imaging recognition. The model adopts a stage-wise heterogeneous convolution module design: FusedMBConv is introduced in the shallow layers to improve training efficiency, while the coordinate attention mechanism is integrated into the deep layers to enhance the capture of pathological features in key regions such as the hippocampus. Additionally, a progressive stochastic depth regularization strategy is introduced to effectively alleviate the overfitting problem in small-sample medical imaging scenarios. Experimental results show that the model achieves excellent performance in the three-class classification task of AD, MCI, and CN, with an early MCI identification accuracy of 98.67%, providing reliable technical support for early clinical intervention. Future research will explore multi-modal data fusion to further improve the diagnostic robustness of the model in complex cases.

CONFLICTS OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGEMENT

This work was supported by the 2025 Annual Undergraduate Research Training Program (No. 2025141), and the 2025 Annual Undergraduate Research Training Program (No. 2025132).

REFERENCES

[1] BRAAK H, BRAAK E. “Neuropathological staging of Alzheimer-related changes”, Acta Neuropathologica, Vol. 82, No. 4, pp. 239-259, 1991. https://doi.org/10.1007/BF00308809

[2] Alzheimer's Association. “2023 Alzheimer's disease facts and figures”, Alzheimer's & Dementia, Vol. 19, No. 4, pp. 1598-1695, 2023. https://doi.org/10.1002/alz.13016

[3] PORSTEINSSON A P, ISAACSON R S, KNOX S, et al. “Diagnosis of early Alzheimer's disease: Clinical practice in 2021”, The Journal of Prevention of Alzheimer's Disease, Vol. 8, No. 3, pp. 371-386, 2021. https://doi.org/10.14283/jpad.2021.23

[4] JACK C R Jr, BERNSTEIN M A, FOX N C, et al. “The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods”, Journal of Magnetic Resonance Imaging, Vol. 27, No. 4, pp. 685-691, 2008. https://doi.org/10.1002/jmri.21078

[5] LITJENS G, KOOI T, BEJNORDI B E, et al. “A survey on deep learning in medical image analysis”, Medical Image Analysis, Vol. 42, pp. 60-88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[6] LIU M, LI F, YAN H, et al. “A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease”, NeuroImage, Vol. 208, p. 116459, 2019. https://doi.org/10.1016/j.neuroimage.2019.116459

[7] KHATRI U, KWON G R. “Diagnosis of Alzheimer's disease via optimized lightweight convolution-attention and structural MRI”, Computers in Biology and Medicine, Vol. 173, p. 108210, 2024. https://doi.org/10.1016/j.compbiomed.2024.108210

[8] PLANT C, TEIPEL S J, OSWALD A, et al. “Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease”, NeuroImage, Vol. 50, No. 1, pp. 162-174, 2009. https://doi.org/10.1016/j.neuroimage.2009.09.046

[9] LIU J, LI M, LUO Y, et al. “Alzheimer's disease detection using depthwise separable convolutional neural networks”, Computer Methods and Programs in Biomedicine, Vol. 203, p. 106032, 2021. https://doi.org/10.1016/j.cmpb.2021.106032

[10] TAN M, LE Q V. “EfficientNet: Rethinking model scaling for convolutional neural networks”, in International Conference on Machine Learning, Long Beach: PMLR, 2019, pp. 6105-6114. http://proceedings.mlr.press/v97/tan19a.html

[11] HU J, SHEN L, SUN G. “Squeeze-and-excitation networks”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City: IEEE, 2018, pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[12] FRISONI G B, FOX N C, JACK C R Jr, et al. “The clinical use of structural MRI in Alzheimer disease”, Nature Reviews Neurology, Vol.6, No. 2, pp. 67-77, 2010. https://doi.org/10.1038/nrneurol.2009.215

[13] HOU Q, ZHOU D, FENG J. “Coordinate attention for efficient mobile network design”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville: IEEE, 2021, pp. 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350

[14] TAN M, LE Q V. “EfficientNetV2: Smaller models and faster training”, in International Conference on Machine Learning, Virtual: PMLR, 2021, pp.10096-10106. http://proceedings.mlr.press/v139/tan21a.html

[15] SANDLER M, HOWARD A, ZHU M, et al. “MobileNetV2: Inverted residuals and linear bottlenecks”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City: IEEE, 2018, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[16] HE K, ZHANG X, REN S, et al. “Deep residual learning for image recognition”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas: IEEE, 2016, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[17] CAO Y, XU J, LIN S, et al. “Global context networks”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 10, pp. 3344-3355, 2019. https://doi.org/10.1109/TPAMI.2019.2952320

[18] YU X, LIU J, LU Y, et al. “Early diagnosis of Alzheimer's disease using a group self-calibrated coordinate attention network based on multimodal MRI”, Scientific Reports, Vol. 14, No. 1, p. 24210, 2024. https://doi.org/10.1038/s41598-024-75143-8

[19] LIN X, LU P, PAN J, et al. “Coordinate attention based 3D-CNN using ghost multi-scale for diagnosing Alzheimer's disease”, in 2024 International Joint Conference on Neural Networks, Yokohama: IEEE, 2024, pp. 1-8. https://doi.org/10.1109/IJCNN60899.2024.10651234

[20] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. “ImageNet classification with deep convolutional neural networks”, in Advances in Neural Information Processing Systems, Lake Tahoe: Curran Associates, Inc.,

2012, Vol. 25, pp. 1097-1105. https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[21] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. “An image is worth 16x16 words: Transformers for image recognition at scale”, in International Conference on Learning Representations, Vienna: ICLR, 2021. https://openreview.net/forum?id=YicbFdNTTy