Integrating Transfer Learning and Attention Mechanisms for Accurate ALS Diagnosis and Cognitive Impairment Detection

Yuqing Xia[1], Jenna M Gregory[2], Fergal M Waldron[2], Holly Spence[2], Marta Vallejo[3]

[1] School of Engineering and Physical Sciences, Heriot-Watt University, UK
[2] Institute of Medical Sciences, University of Aberdeen, Aberdeen, AB25 2ZD, UK
[3] School of Mathematical and Computer Sciences, Heriot-Watt University, UK

September 22, 2024

Abstract

Amyotrophic lateral sclerosis (ALS) is a fatal neurological disease characterized by motor deterioration and cognitive decline, leading to respiratory failure. Early diagnosis is crucial but challenging due to the undefined risk population and the complexity of sporadic ALS. In this study, we used a dataset of 190 autopsy brain images from the Gregory Laboratory at the University of Aberdeen to develop Miniset-DenseSENet, a convolutional neural network combining DenseNet121 with a Squeeze-and-Excitation (SE) attention mechanism. Our model not only distinguishes ALS patients from control groups with 97.37% accuracy but also detects cognitive impairments, which are increasingly recognized as a critical but underdiagnosed feature of ALS. Miniset-DenseSENet outperformed other transfer learning models, achieving a sensitivity of 1 and specificity of 0.95. These findings suggest that integrating transfer learning and attention mechanisms into neuroimaging analysis could enhance clinical diagnostic capabilities, enabling earlier and more accurate diagnosis of ALS and cognitive impairment. This approach has the potential to improve patient stratification, guide clinical decision-making, and inform the development of personalized therapeutic strategies.

Keywords: Amyotrophic lateral sclerosis, TDP-43 protein, Cognitive impairment, Transfer learning, Attention mechanisms

1 Introduction

Amyotrophic lateral sclerosis (ALS) is a condition that progressively degenerates the neurons responsible for motor control [1]. As a patient's motor neurons degenerate and die, the patient's brain loses the ability to initiate and control voluntary movement and eventually dies from respiratory failure [2]. ALS has a median survival of only 2–4

years. The ALS Therapy Development Institute projects a global rise in ALS instances from 222,801 in 2015 to 376,674 by 2040, marking a 69% surge [1]. Moreover, the average delay from the first appearance of symptoms to the final diagnosis is about 14 months [3].

In an analysis of autopsy cases from ALS patients, a specific protein called the TAR DNA-binding protein 43 (TDP-43) was found to accumulate abnormally in the cytoplasm [4]. This abnormal accumulation is observed in the motor neurons of nearly all patients with sporadic ALS and in the majority of patients with familial ALS ($> 84\%$) [5]. The frequent presence of TDP-43 in almost all sALS patients and most fALS patients suggests that it may play a key role in the ALS disease process [2]. Additionally, Gregory *et al.* found that although all ALS patients exhibit TDP-43 pathology in extramotor brain regions, only some present with cognitive impairment [4]. The progression and ultimate outcome of ALS vary significantly between patients, making it difficult to accurately predict the disease's course [6, 7].

Deep learning models have achieved unexpected progress in classification and recognition tasks. For example, convolutional neural networks (CNNs) process images by mimicking the workings of the human visual system [8], while generative adversarial networks (GANs) are very effective in tasks such as image generation, editing, and enhancement [9]. Given its ability to recognize complex and subtle patterns and relationships, deep learning has been widely used in medical image analysis and image processing in neuroimaging research with varying degrees of success. However, medical data is often difficult and costly to obtain, often requiring invasive procedures (e.g., biopsies and spinal fluid extractions), uncomfortable tests (e.g., blood tests), or time-consuming procedures (e.g., magnetic resonance imaging) [10]. When deep learning is applied to the medical field, the problem of data sparsity becomes particularly prominent. For example, Berg *et al.* used deep learning to combine MRI data with clinical features of only 135 ALS patients to predict the survival rate of ALS patients [11]. Similarly, Liu *et al.* used only T2-weighted structural MRI data from 46 patients to predict whether they had ALS [12] and Thome *et al.* conducted an ALS classification study using unimodal and multimodal random forest classifiers, using data provided by 97 ALS patients and 59 healthy control participants [13]. These studies show that data sparsity and difficulty obtaining data remain major challenges in ALS.

Additionally, capturing complex disease features in small datasets is challenging. For example, Khuzaim Alzahrani *et al.* used an integrated multilevel encoder-decoder architecture that combines U-shaped networks and GANs to predict ALS disease and its progression [14]. However, this method requires a large amount of high-quality annotated data to train the models. Rahul Atmaramani et al. employed self-supervised learning to process high-content imaging data to predict cellular morphological changes associated with ALS [15]. Although this method reduces the need for large-scale annotated data, the quality of feature learning in self-supervised learning depends on the design of pretext tasks, which may struggle to capture the most relevant features for complex disease characteristics. Despite the advantages of these methods, they still face many challenges in practical applications.

Therefore, developing robust models that can effectively handle small datasets and capture complex disease characteristics remains an important research direction. Data

augmentation and transfer learning are two widely adopted strategies to alleviate the limitations of small datasets [16].

Data augmentation involves creating new training samples by applying various transformations to the original data. This is particularly useful in increasing the size and variability of the training dataset [17]. Data augmentation methods involve rotating or reflecting the original image, scaling it up or down, and translating it to create new images. For example, by augmenting the training dataset with rotated and flipped versions of medical images, the model can become more robust to changes in image orientation and position, ultimately enhancing its generalization ability [18]. In the field of medical image processing, data augmentation has been widely used. For example, in tumor detection and classification in radiological images, augmenting the dataset with various transformations has been shown to significantly improve model performance [19]. Similarly, in brain MRI analysis, augmenting the dataset with elastic deformation helps the model better capture the variability of anatomical structures [20].

Transfer learning is another effective method to address the issue of limited training data. Using models pre-trained on large datasets such as ImageNet allows transfer learning to transfer the learned features to a new, smaller dataset [21]. This process involves fine-tuning the pre-trained model on a specific medical dataset, which can improve the model's initial performance and reduce the need for a large amount of labeled data [22]. For example, pre-trained CNNs such as ResNet [23], VGG [24], and DenseNet [25] have been successfully applied to tasks like diabetic retinopathy detection [26], breast cancer classification [27], and brain tumor segmentation [28]. We use DenseNet121 as our transfer learning model because DenseNet has efficient feature reuse and improved gradient flow. Each layer in DenseNet is directly connected to every previous layer, which promotes the retention and transfer of information, helping to better capture complex patterns and details [25].

However, transfer learning alone may not fully address the complexity of disease features in medical images. To further enhance model performance, we use attention mechanisms to optimize feature extraction and highlight the most relevant parts of the input data. Attention mechanisms can be intuitively explained using the human visual system [29], and they have good integration properties with many methods being very lightweight, such as the squeeze-and-excitation (SE) model and the convolutional block attention module (CBAM) model. This makes them well-suited for combination with deep learning [30]. Generally, attention mechanisms determine which information is important by assigning weights. This process improves model accuracy and provides better interpretability of the results. Specifically, the SE module's channel-level attention mechanism can better capture and utilize fine-grained features within DenseNet, enhancing the model's recognition and classification performance.

Due to the unclear definition of the risk group for sporadic ALS, it is challenging to find early, pre-symptomatic diagnostic biomarkers. Therefore, it is very difficult to obtain a large dataset. In this project, we used a dataset of only 190 post-mortem brain images provided by the Gregory Laboratory at the University of Aberdeen. The images are stained with TDP-43 RNA aptamer (TDP-43APT) [31, 32], using the Standard Operating Procedure published previously [33].

In this study, we exploit the combination of transfer learning and attention mechanism to develop an effective model framework named Miniset-DenseSENet. The model leverages the powerful feature extraction capabilities of pre-trained DenseNet121 and combines SE modules to enhance feature representation. By focusing on the most relevant features in the dataset, our approach aims to improve the diagnosis and understanding of ALS. Our main contributions are as follows:

1. We explored the differences in postmortem brain images between individuals with and without ALS, as well as ALS patients with and without cognitive impairment.
2. We developed a novel deep learning model framework named Miniset-DenseSENet, which integrates DenseNet121 and SE modules. This framework leverages transfer learning to enhance feature extraction and incorporates attention mechanisms to improve the quality of feature representations.
3. We performed a comprehensive evaluation of various transfer learning models and their attention mechanism variants on the ALS classification problem. Our experiments included a wide range of models such as baseline CNN, ResNet18, DenseNet121, ResNet18+SE, ResNet18+CBAM, and Miniset-DenseSENet, providing a detailed performance comparison and identifying the most effective approaches.
4. We provided visual explanations for the model's classification decisions using gradient-weighted class activation mapping (Grad-CAM). These visualizations help to interpret the model's behavior and highlight the critical regions in brain images that influence the classification results, thereby enhancing the transparency and trustworthiness of the model.

The remainder of this paper is structured as follows: Section 2 reviews related work on Miniset-DenseSENet. Section 3 covers data preprocessing methods, the structural design of Miniset-DenseSENet, and the experimental evaluation methods using traditional CNNs as a benchmark. Section 4 presents a comparison of Miniset-DenseSENet with other transfer learning and attention mechanism models, alongside experimental results compared to basic CNNs. In Section 5, various issues and data phenomena are discussed. Finally, Section 6 provides the conclusions.

2 Background

The solution proposed in this paper combines transfer learning DenseNet121 and the SE module. In this section, we will introduce some background knowledge of these techniques.

2.1 The DenseNet121 Model

Densely Connected Convolutional Networks (DenseNet) are characterized by dense connectivity, where each layer is directly connected to every other layer [25]. This highly parameter-efficient architecture promotes feature transfer and reuse and mitigates the vanishing gradient problem.

The core strength of DenseNet121 lies in its feature reuse mechanism, which makes the model easy to train and highly parameter-efficient. The relationship is described by the following formula:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{1}$$

Where x_l represents the output of the l-th layer, while H_l is a composite function composed of convolution, batch normalization, and ReLU activation function. The notation $[x_0, x_1, ..., x_{l-1}]$ denotes the concatenation of all feature maps from layer 0 to layer $l-1$ along the channel dimension.

The architecture of DenseNet121 is illustrated in Figure 1(3). Each layer in this structure utilizes the information from all preceding layers, resulting in high parameter efficiency and effective feature utilization. Compared to traditional network architectures, DenseNet121 exhibits a strong regularization effect, which helps reduce the risk of overfitting, particularly when training with fewer samples.

2.2 The Squeeze-and-Excitation Module

The SE module, introduced by Hu *et al.* in 2018 [34], employs channel attention mechanisms and can be added to any baseline architecture. This mechanism facilitates feature recalibration, enabling the network to use global information to selectively enhance informative features while suppressing the less useful ones [29]. Figure 1(1) illustrates the fundamental structure of an SE block. The structure of the SE module can be divided into three core parts: transformation, global information embedding (squeeze) and adaptive recalibration (excitation).

In the transformation stage, the SE module extracts features of each channel from the input feature map X through the convolution operation v_c and sums them up to form a new feature vector u_c for channel c. The formula is as follows:

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{2}$$

Where u_c represents the result of the feature extraction of the cth channel, X is the input feature map, v_c^s is the weight of the convolution kernel at the sth position, and x^s is the value of the feature map at the same position. This operation ensures that the features of each channel can be precisely captured before entering the squeeze stage. In the squeeze stage, the SE module compresses global spatial information into channel descriptors through a global average pooling operation. The formula is as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{3}$$

Where z_c is the global feature of channel c, H and W are the height and width of the feature map, respectively. In the excitation stage, to fully capture channel-wise dependencies, the SE module must be able to learn non-linear interactions between channels and non-mutually exclusive relationships. Therefore, the SE module adopts a simple gating mechanism with a Sigmoid activation. The formula is as follows:

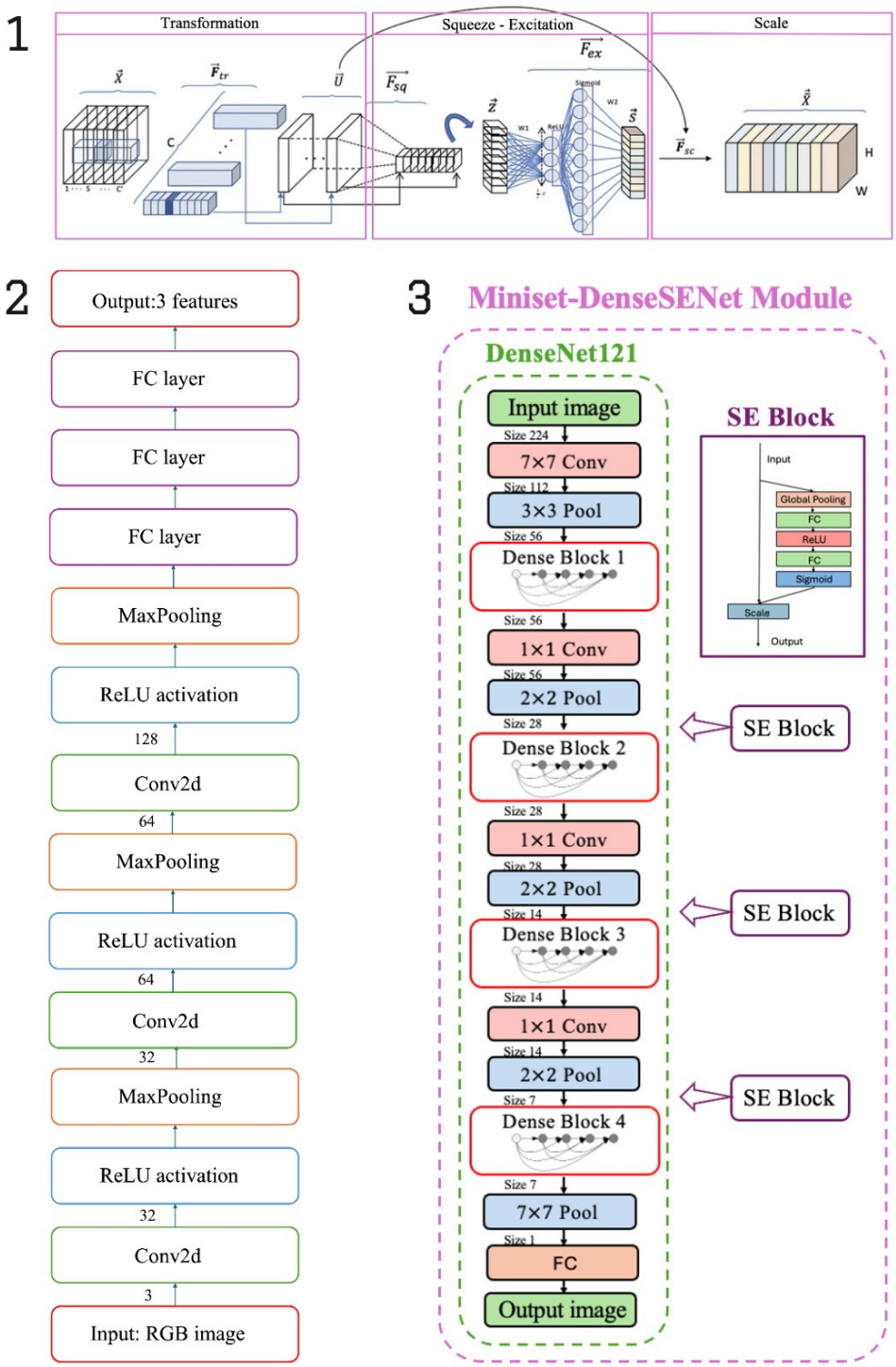


Fig. 1: (1) The SE block enhances network performance by adaptively recalibrating channel-wise feature responses. It first uses global average pooling to "squeeze" spatial information into a channel descriptor, followed by an "excitation" phase with fully connected layers that generate weights for each channel. These weights are applied to the original feature map to emphasize important features and suppress less relevant ones, improving the model's representational capacity with minimal computational cost. (2) The model architecture of a basic CNN. (3) The model architecture of DenseNet121 and its model architecture after integrating the SE module.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) \tag{4}$$

Where z is the global feature vector obtained through the Squeeze operation, W is the learned parameter, and σ represents the Sigmoid activation function, g is the fully connected layer operation. The final output of this block is obtained by rescaling the transformed output U using activations as follows:

$$\tilde{x}_c = F_{\text{scale}}(u_c, \delta_c) = \delta_c \cdot u_c, \tag{5}$$

Here, $\tilde{X}$ is the set of recalibrated feature composed of $\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_C\}$, and $F_{\text{scale}}(u_c, \delta_c)$ refers to the channel-wise multiplication between the feature map $u_c \in \mathbb{R}^{H \times W}$ and the scalar δ_c.

3 Methodology

3.1 Image Dataset

In this project, we apply deep learning techniques to classify a dataset of 190 images provided by Gregory's lab at Aberdeen University. The dataset comprises three categories: 60 images in the concordant group, which are images of patients with TDP-43 pathology in the extramotor cortex and cognitive impairment; 60 images in the discordant group, which are images of patients with TDP-43 pathology in the extramotor cortex but without cognitive impairment. Finally, the control group consists of 70 images from individuals without ALS, characterized by the absence of TDP-43 pathology in the extramotor cortex (Figure 2).

The brain tissue images were created using immunohistochemistry techniques, allowing for the precise staining and visualization of TDP-43 pathology in specific cortical areas. Immunohistochemistry enables the detection of specific proteins within the brain tissue samples, providing a detailed view of the presence and distribution of TDP-43 inclusions that are critical for understanding ALS pathology.

The brain images in the dataset are taken from two key regions: Brodmann area 44 (BA44) and Brodmann area 46 (BA46). BA44, located in the left inferior frontal gyrus and part of Broca's area, is primarily associated with language production and higher-order cognitive functions [35]. BA46, situated in the dorsolateral prefrontal cortex, is involved in executive functions such as working memory, cognitive control, and decision-making [36]. These regions are of particular interest due to their involvement in both cognitive processing and motor control, making them critical areas for studying the neuroanatomical basis of motor and cognitive impairments in neurodegenerative diseases.

The inclusion of samples from these two key brain regions enables the dataset to provide a more detailed investigation into the relationship between motor and cognitive dysfunction. This focus on BA44 and BA46 helps bridge the gap between cognitive decline and motor dysfunction, facilitating a deeper understanding of the distinct patterns observed in concordant and discordant cases.

All post-mortem tissue was collected with ethics approval from the East of Scotland Research Ethics Service (16/ES/0084) in line with the Human Tissue (Scotland)

Act (2006). The use of post-mortem tissue for studies was reviewed and approved by the Edinburgh Brain Bank ethics committee and the Academic and Clinical Central Office for Research and Development (ACCORD) medical research ethics committee (AMREC). Clinical data were collected as part of the Scottish Motor Neurone Disease Register (SMNDR) and Care Audit Research and Evaluation for Motor Neurone Disease (CARE-MND) platform, with ethics approval from Scotland A Research Ethics Committee (10/MRE00/78 and 15/SS/0216) and have been published previously [4]).

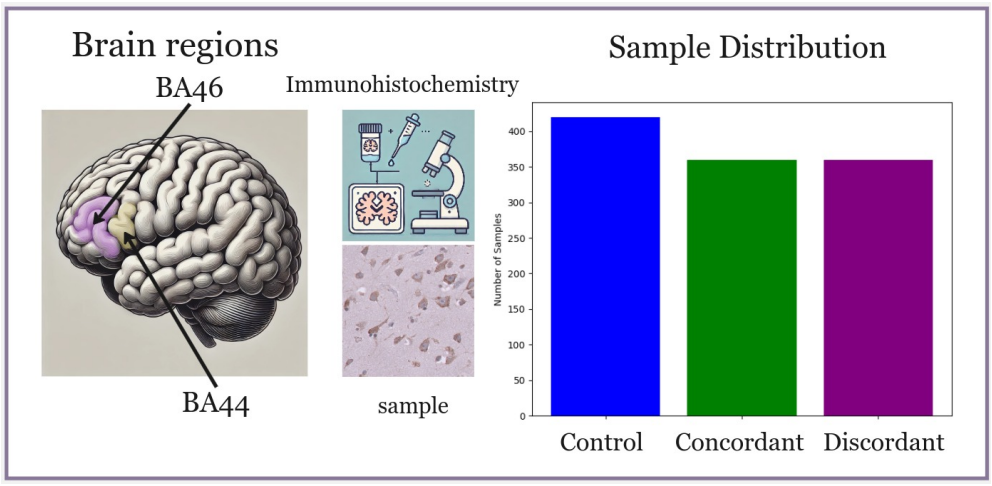


Fig. 2: The figure provides a comprehensive overview of the dataset and its neuroanatomical focus. On the left, a brain diagram highlights the regions BA44 and BA46, which are the areas from which the brain tissue samples were taken. In the center, the top section features an icon representing the immunohistochemistry techniques used to create the brain tissue images, while below it is a sample image from the dataset. On the right, a bar chart shows the sample size distribution across three classes: control, concordant, and discordant. Concordant samples represent individuals showing signs of cognitive decline, while discordant samples display motor dysfunction without cognitive impairment. Together, these visual elements emphasize the neuroanatomical regions, image acquisition techniques, and the distribution of data across key categories in the study.

3.2 Data Augmentation

Training deep CNNs on small datasets and enhancing the generalization capabilities of these models is an extremely challenging task. During training, models with poor generalization tend to overfit the training data. Data augmentation is an effective strategy to solve this problem. It minimizes overfitting problems by extending the data representation to include a more comprehensive set of potential data points [17].

This project uses nine data augmentation methods, including geometric transformation and various image processing functions, to achieve comprehensive image

augmentation. Geometric transformation mainly changes the spatial layout of image content, while non-geometric transformation involves adjusting the appearance of the image, such as color, brightness, contrast, etc. The specific methods are as follows:

- Flipping: Flip the image horizontally.
- Rotating: Rotate the image 90 degrees.
- Image Cropping: Cut off a specified border width evenly from all sides of each original image to obtain a smaller central area than the original image.
- Image Scaling: Adjust the size of the image.
- Perspective Transformation: Simulates the visual effects of observing an image from different angles.
- Grayscale Conversion: Convert color images to grayscale images.
- Grayscale Image Denoising: Reduce random noise in grayscale images.
- Color Image Denoising: Reduce random noise in color images.
- Color Image Denoising: Reduce random noise in color images.
- Color Image Denoising and Enhancing Image Quality: Bilateral filtering is used to reduce noise in color images, and a binary mask is created for edge protection through the Canny edge detection [37] and expansion operations. Finally, an adaptive histogram equalization method is applied to the luminance channel of the image to improve the contrast and sharpness of the image.
- Add noise: Add random noise to the image to improve the robustness of the model.
- Brightness Enhancement: Improve the overall brightness of the image.

3.3 Model Architecture: DenseNet121+SE

In this project, we combine DenseNet121 with the SE module. The structure diagram is shown in Figure 1(3). In the DenseNet architecture, layers within each dense block are connected in a feed-forward manner to all other layers, allowing the network to deeply extract and fuse features from previous layers and increase the diversity of new features [25].

A key challenge in this feature fusion process is managing feature redundancy. The SE module addresses this by automating feature learning, enabling DenseNet121+SE to suppress unnecessary features and minimize the impact of redundancy. Additionally, placing SE modules between dense blocks, rather than within them, helps avoid the risk of overfitting; SE layers within dense blocks may overlearn feature weights while still exploring new features [38]. In contrast, an SE layer positioned between dense blocks can enhance useful ALS-related features and suppress redundant ones after feature extraction.

Through this combination, we aim to achieve more accurate and efficient performance in ALS image classification tasks.

3.4 Baselines

This section describes the baseline model used in experiments to benchmark the ALS classification system. In particular, we consider a well-known alternative for image classification: CNN. CNN surpasses other machine learning methods in terms of predictive power and has been applied in a variety of image processing tasks, such as handwriting recognition and image recognition problems [39]. Therefore, we selected the results of a CNN model with three convolutional layers as a baseline for comparison, as shown in the Figure 1(2).

3.5 Performance - Evaluation Metrics

In this project, model performance was quantified by assessing a suite of metrics, with each trained model subjected to five independent tests. Evaluative criteria included the Matthews Correlation Coefficient (MCC) [40], accuracy, sensitivity, and specificity.

MCC is a performance indicator for classification problems based on the correlation coefficient between the observed and predicted classifications. Its values range from -1 to 1, where 0 represents a random guess, 1 a perfect prediction, and -1 represents a completely inconsistent prediction.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

Accuracy represents the proportion of images correctly identified out of the overall number of predictions made.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Specificity refers to the accuracy of a test in correctly recognizing individuals who are not afflicted with the disease.

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

Sensitivity measures a test's capacity to detect individuals affected by the disease accurately.

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

In these formulas, TP, TN, FP, and FN stand for True Positives, True Negatives, False Positives, and False Negatives, respectively.

3.6 Visualization Techniques

CNN is usually regarded as a black box model. To improve its interpretability, this study adopts the Grad-CAM technology [41]. The technique provides the model with a "visual explanation" of the decision by highlighting the areas of the image that CNN focused on when making a specific decision.

This technique relies on the gradient information of the model to generate a heatmap by analyzing the impact of specific categories on the output of the last convolutional layer. This process can be simplified by the following formula:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{10}$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{11}$$

Where α_k^c denotes the importance weights for each channel in the feature map A^k with respect to class c. $L_{\text{Grad-CAM}}^c$ is the class-discriminative localization map for class c, which is a matrix of width u and height v. Before softmax transformation, y^c denotes the score of category c, and gradient $\frac{\partial y^c}{\partial A^k}$ denotes the effect on the category score y^c by measuring the change in feature map activation A^k.

This heat map reveals which parts of the image the model values when identifying or classifying a specific class, indicating which regions contribute most to the ultimate decision. By applying Grad-CAM, researchers can visually see which part of the tissues plays a decisive role in diagnostic decisions when the model analyzes postmortem brain images of ALS patients and controls.

4 Experimental Results

In this section, a quality inspection is first performed on the 190 images to evaluate the impact of image quality on model classification performance. Subsequently, the impact of using various data augmentation methods on model classification performance is explored. Next, the performance of Miniset-DenseSENet on image classification problems was tested and compared with four other models: ResNet18, DenseNet121, ResNet18+SE, ResNet18+CBAM, and the baseline model CNN. Finally, the Grad-CAM technique is utilized to examine the focus of the model in image classification tasks.

4.1 Original Image Quality Analysis

Here, 190 images are quality-checked to ensure all training images are clear and valid. At the same time, quality inspection of the data-enhanced images can ensure the effectiveness of data enhancement. The testing of the Laplacian variance, brightness average, contrast (expressed as standard deviation), signal-to-noise ratio (SNR) and other indicators of each image allows us to analyze the quality of the dataset, following standard practices in image quality assessment [42]. In addition, we chose violin plots to show the distribution and density of these indicators, which helps us understand the fluctuations in image quality in the dataset more intuitively.

The detection results of the images are shown in Figure 3. Observing these images, it was found that the mean distribution of contrast and brightness is relatively concentrated, and there are no extreme outliers, indicating that most images are relatively consistent in brightness and contrast. The SNR distribution is narrow, which indicates

that the images are relatively balanced in signal quality without excessive noise. However, the distribution of the Laplacian variance shows that there are varying degrees of image sharpness but no outliers. The images with the minimum and maximum Laplacian variance are shown in Figure 3(1) and Figure 3(2). It can be observed that most images are acceptable in terms of clarity. These distribution plots indicate that the image quality of the dataset is appropriate and can be used for further data enhancement and deep learning training. See A.1 for further details.

4.2 Data Augmentation

This section uses nine methods to perform data augmentation on the dataset. The approaches include flipping, rotating, grayscale conversion, grayscale image denoising, color image denoising, image scaling, image cropping, adding noise, and perspective transformation. Taking the original image in Figure 4 as an example, the extended images show the resulting images processed by different data augmentation methods.

4.3 Experimental Setup

In the model training phase of this study, Python 3.8 was selected as the programming language to ensure efficient development and model implementation. In this study, the performance of Miniset-DenseSENet was compared with that of four other models: ResNet18, DenseNet121, ResNet18+SE, and ResNet18+CBAM, as well as with the baseline model CNN. The training dataset used in this experiment includes the original data set and data processed by nine different data augmentation techniques, with a total of 1140 images. However, given the limited scale of the available training dataset, all CNN modules in this study are pre-trained on ImageNet. When using these pre-trained models, most of the layers are frozen, with only zero to two modules in a trainable state, while the Global Average Pooling (GAP) layer and Fully Connected (FC) layer are set to trainable states for fine-tuning specific tasks. Here are the training strategy and freezing details:

- ResNet18: All modules of the pre-trained ResNet18 model are frozen, and only the GAP and FC are set as trainable layers.
- ResNet18+SE: All modules of the pre-trained ResNet18 model are frozen, and the GAP, FC and SE modules are set as trainable layers.
- ResNet18+CBAM: Most modules of the pre-trained ResNet18 model are frozen, and only the last two modules, the GAP, the FC and the CBAM modules, are set as trainable layers.
- DenseNet121: All modules of the pre-trained DenseNet121 model are frozen. The GAP and the FC are set as trainable layers.
- Miniset-DenseSENet: All modules of the pre-trained DenseNet121 model are frozen, and the GAP, FC and SE modules are set as trainable layers.

The selection of hyperparameters for each model was informed by careful consideration of various factors, including limitations imposed by GPU memory, training efficiency, the risk of overfitting, and the inherent size of the models. Detailed configurations of the hyperparameters are systematically presented in Table 1 and Table 2.

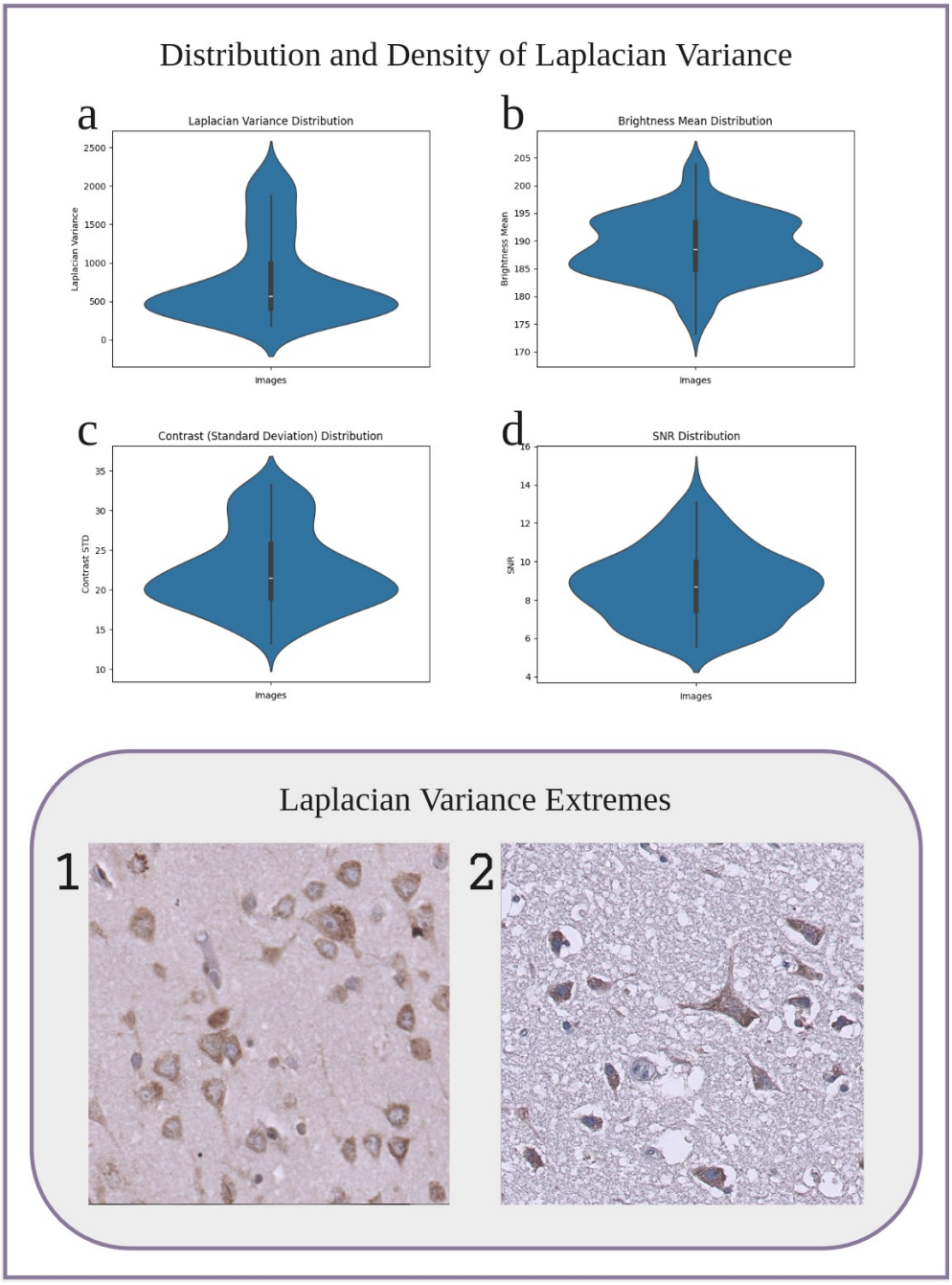


Fig. 3: Top: Violin plots showing the distribution and density of four image features: (a) Laplacian variance, (b) signal-to-noise ratio (SNR), (c) brightness mean, and (d) contrast. These plots illustrate the variability and spread of these metrics across the dataset. Bottom: Two ALS image samples representing the extremes of Laplacian variance. Image 1 shows the minimum variance, while Image 2 shows the maximum variance, highlighting the visual differences in image sharpness between the two samples.

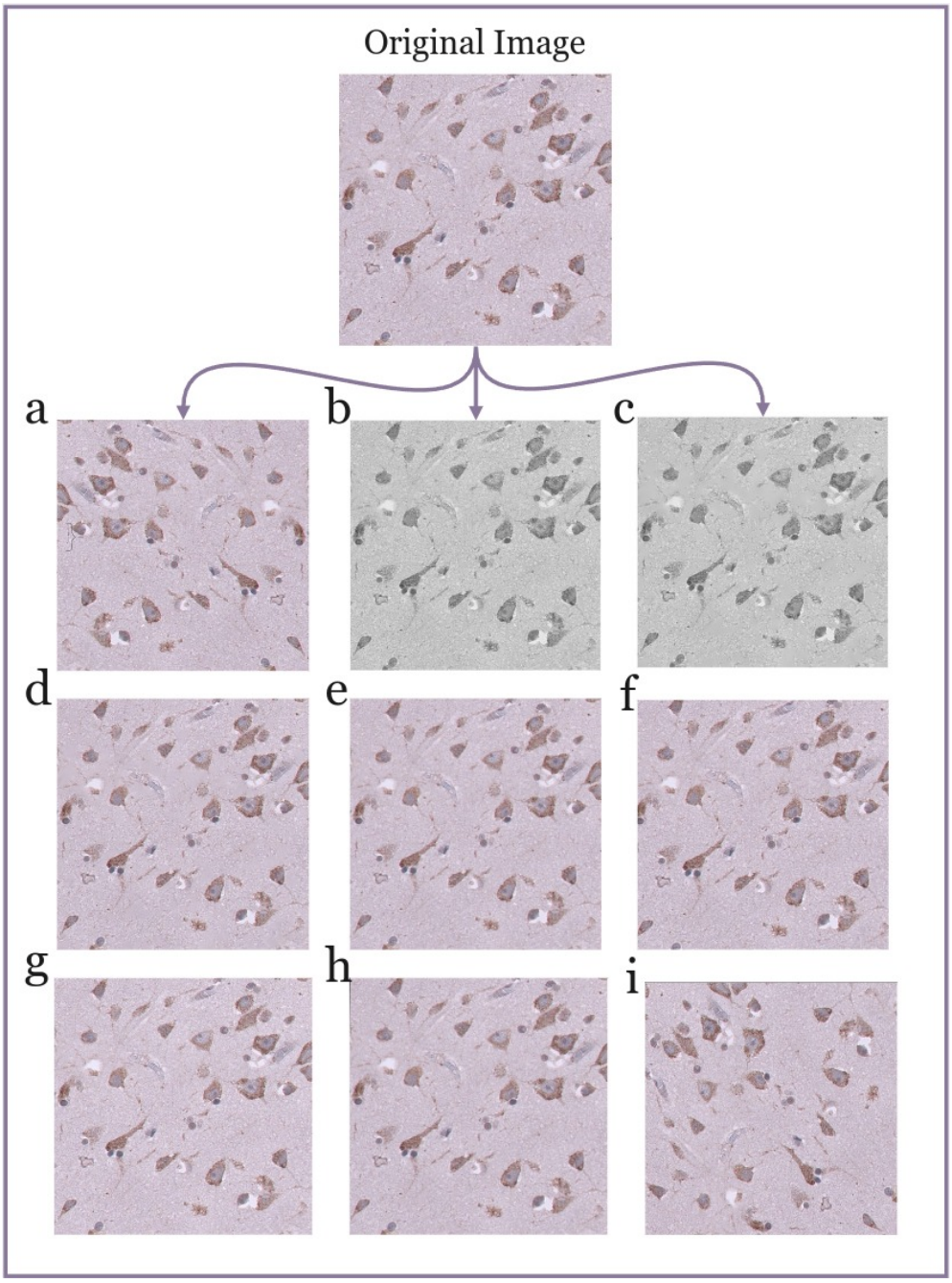


Fig. 4: The figure above displays the original image (top), followed by nine images augmented using various techniques. The augmented images, arranged in a 3x3 matrix, are as follows: (a) Flipping, where the image is mirrored horizontally; (b) Grayscale, converting the image to grayscale; (c) GrayDenoised, applying noise reduction to the grayscale image; (d) ColorDenoising, applying noise reduction while maintaining the original color scheme; (e) Scaling, resizing the image without altering the aspect ratio; (f) Cropping, selecting and displaying only a portion of the original image; (g) Adding Noise, introducing random noise to the image; (h) Perspective Transformation, altering the perspective view of the image; and (i) Rotating, rotating the image by a specified angle. These augmentations highlight the variety of transformations applied to the original data to enhance model robustness.

Table 1: The applied hyperparameters of different CNN models

HPs / CNNs	Learning rate	Batch size	Optimizer	Epochs	Weight decay
ResNet18	0.0001	64	Adam	150	0.008
DenseNet121	0.00008	32	Adam	250	0.08
ResNet18+SE	0.0005	128	Adam	300	0.0008
ResNet18+CBAM	0.0001	128	Adam	200	0.008
Miniset-DenseSENet	0.0001	32	Adam	200	0.001

Table 2: The applied hyperparameters of the basic CNN

Hyperparameter	Learning rate	Batch size	Dropout	Optimizer	Epochs	Patience
Value	0.0001	64	0.3	Adam	300	5

Table 3: Experimental settings: dataset

Dataset	Training Size	Validation Size	Test Size
190	114	38	38
After Data argumentation	9×114	0	0
Total	1140	38	38

4.4 Model Performance Comparison

In this section, we compare the performance of Miniset-DenseSENet with five other models: ResNet18, DenseNet121, ResNet18+SE, ResNet18+CBAM, and the baseline model CNN. We evaluate their performance by analyzing the MCC, computational complexity, parameter efficiency, accuracy, and other indicators of the models. Additionally, we performed a visual analysis of the correct and incorrect results of model classification.

4.4.1 Dataset Division

The 190 image datasets are divided into training set, validation set and test set according to the distribution ratio of 6:2:2. The training dataset in this experiment was processed using nine different data augmentation techniques delineated in subsection 4.2., and there were a total of 1,140 images in the final training dataset. The specific data is shown in Table 3. Figure 2(right) illustrates the categorical distribution within the training dataset, indicating an absence of imbalance across the classes. Sample imbalance can significantly bias the model by causing it to favor the majority class, which often results in reduced generalization and poorer performance on minority classes [43].

4.5 Performance Evaluation over different models

In this subsection, we undertook a detailed training and evaluation of Miniset-DenseSENet and four other distinct models: ResNet18, DenseNet121, ResNet18+SE, and ResNet18+CBAM. These models were developed based on the foundational ResNet and DenseNet architectures, with specific enhancements through integrating SE and CBAM attention mechanisms in designated variants. These models were juxtaposed against a traditional CNN baseline to benchmark their performance.

We performed five independent runs on the trained models. The number of iterations (epochs) of model training was set between 150 and 300. Figure 5(top) shows the trend of model accuracy as the number of training iterations increases, where the accuracy shown is the median value among the five runs. Subsequently, the accuracy outcomes are presented as a box plot in Figure 5(middle, left). A comparative analysis of these figures reveals that the accuracy of the five transfer learning models is significantly better than that of the conventional CNN approach. Notably, the accuracy of the DenseNet model augmented with the SE module consistently exceeds 90%, with an optimal performance of 97.37%.

Further, we employed the MCC, sensitivity, and specificity metrics to conduct a performance evaluation of the models, as shown in Figure 5(middle, right), Figure 5(bottom, left) and Figure 5(bottom, right). The results showed that the control group performed well in terms of sensitivity and specificity. Specifically, in five experiments, the specificity of the control group for ResNet18, ResNet18+SE, and ResNet18+CBAM models reached 1, and the sensitivity for ResNet18 and Miniset-DenseSENet achieved a score of one in four out of five measurements. However, the sensitivity and specificity performance for the Concordance and Discordance classes did not meet the high standards observed in the control group and raised some stability concerns. Miniset-DenseSENet demonstrated the most superior results among them. The average sensitivity for both classes was approximately 80%, and the specificity was around 93%. In the Appendix, Figure 7 displays the confusion matrices for the six models, further validating these observations.

A comparative analysis of the six models, as observed in Table 4, shows higher computational complexity, measured in FLOPs, is generally associated with increased accuracy. However, this relationship is not linear, indicating that simply increasing computing resources does not guarantee a proportional improvement in performance [44]. The DenseNet architecture utilizing feature reuse exhibits higher parameter efficiency compared to ResNet's residual learning method. Integrating attention mechanisms such as SE and CBAM into the traditional architectures produces mixed results. The specific reasons will be analyzed in the discussion. Miniset-DenseSENet achieves the highest accuracy among the evaluated models, indicating a positive synergy between the DenseNet architecture and the SE mechanism. In contrast, the accuracy of ResNet18+CBAM drops emphasizes that the effectiveness of the attention mechanism depends on the underlying model architecture and the specific task.

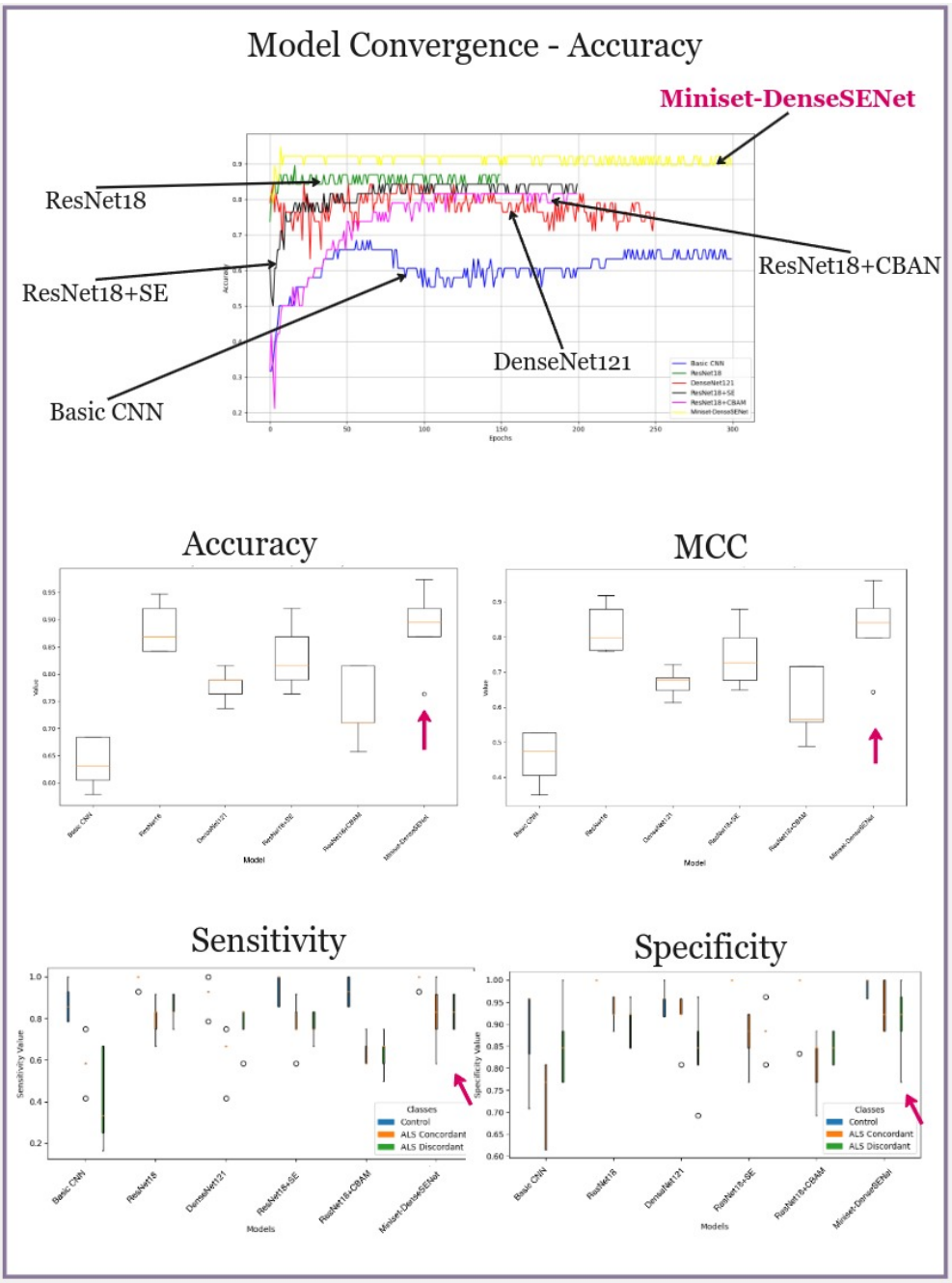


Fig. 5: The figure illustrates the performance comparison between different models, arranged from top to bottom. At the top, the graph displays the accuracy values collected over consecutive epochs for six models: Miniset-DenseSENet, ResNet18, ResNet18+SE, ResNet18+CBAM, DenseNet121, and Basic CNN. The models are listed in order of performance from best to worst accuracy convergence. In the middle row, boxplots show the accuracy and Matthews correlation coefficient (MCC) over five repetitions, where Miniset-DenseSENet demonstrates the highest accuracy and MCC, followed by ResNet18. The final row presents sensitivity and specificity values for each model, with results and variance aligning with the accuracy and MCC observations from previous rows. These plots confirm the superior performance and consistency of Miniset-DenseSENet compared to the other models.

Table 4: Comparison of model parameters, FLOPs and average accuracy

Models	FLOPs	Params	AvgAccuracy
Basic CNN	1.9G	95232	63.16%
ResNet18	5.9G	11.2M	86.84%
DenseNet121	9.2G	7.0M	78.95%
ResNet18+SE	5.9G	11.2M	81.58%
ResNet18+CBAM	5.9G	11.22M	71.05%
Miniset-DenseSENet	9.2G	7.1M	89.47%

5 Grad-CAM Analysis of Correct and Incorrect Cases

In this section, we use Grad-CAM to visualize the classification rationale of the Miniset-DenseSENet trained model, exploring instances of both correct and incorrect classifications through visualization, as shown in Figure 6. The principle of creating this diagram is similar to that of confusion matrices. The horizontal axis represents the predicted categories of the images, while the vertical axis represents the true categories. From the distribution of the heatmap, it can be seen that the model correctly recognizes images along the diagonal, while misclassifications are found at other positions.

6 Discussion and Analysis

In this project, our decision not to employ data augmentation methods, such as colour denoising and enhancement or brightness adjustments, was motivated by the need to preserve the integrity of critical biomarkers in postmortem brain images of ALS patients. In neurodegenerative diseases, maintaining the fidelity of pathological signals is essential, especially when the underlying features of interest may be subtle or obscured by extraneous enhancements. Excessive manipulation of image characteristics can inadvertently introduce biases, causing models to focus on irrelevant or non-specific areas, thus diverging from the true disease-related features that are key to accurate diagnosis and mechanistic understanding. For ALS, where the presence of TDP-43 proteinopathy is central to the disease, ensuring that machine learning models identify and prioritise relevant features is essential.

The integration of advanced techniques such as transfer learning and attention mechanisms has significantly improved model performance compared to traditional CNN training methods (see Figure 5). Specifically, our results show that ResNet18 outperforms DenseNet121, which is noteworthy given that DenseNet models are often favoured for their deep, densely connected architectures. However, introducing the SE module into our Miniset-DenseSENet architecture shifted this balance, enabling superior performance compared to both ResNet18 and ResNet18+SE. This observation is particularly relevant in the context of machine learning research in ALS, where data is often limited. Models like ResNet18, with fewer parameters, are advantageous when data availability is restricted, as they mitigate the risk of overfitting while still capturing disease-relevant features. However, as demonstrated by the enhanced performance

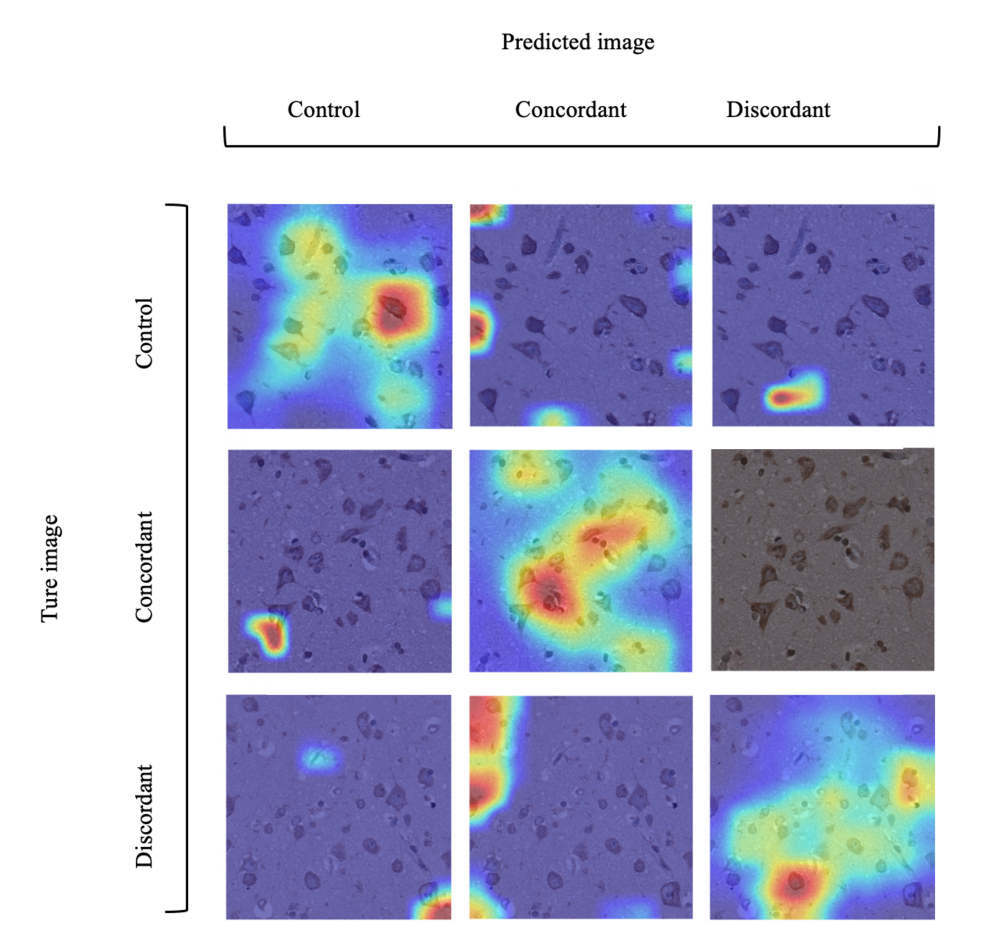


Fig. 6: Classification Performance Analysis with Miniset-DenseSENet: Correct and Misclassified Instances

of Miniset-DenseSENet, incorporating attention mechanisms through the SE module can selectively amplify critical input features in dense layers, improving the model's capacity to generalise despite the small dataset.

The machine learning community's ongoing work developing and refining models such as these is crucial for advancing our understanding of ALS. The pathophysiological mechanisms of ALS, particularly those involving TDP-43, are highly complex and not fully understood. Machine learning offers a unique lens through which these mechanisms can be explored, as models trained on large datasets can identify patterns and associations that may be difficult to detect through conventional analytical approaches. The integration of attention mechanisms, in particular, has shown promise in emphasising subtle but meaningful features likely to be associated with pathological changes.

However, it is important to recognise that the effectiveness of attention mechanisms is contingent upon several factors, including the complexity of the model, the nature of the dataset, and the optimisation strategy. In cases where simpler models like ResNet18 can effectively capture relevant features, further complexity may be redundant and detract from performance. This insight highlights the importance of tailoring model architecture to the specific requirements of a given task, especially in scenarios involving rare diseases like ALS, where data scarcity is often a limiting factor. The superior performance of the DenseNet121+SE model compared to ResNet18+SE may reflect the inherent compatibility of the DenseNet architecture with the SE module in this application. DenseNet's dense connections facilitate the propagation of information across layers, which, when combined with the feature recalibration capability of the SE module, enhances the model's ability to focus on disease-relevant areas in a more holistic manner.

This work underscores the critical role that the machine learning community plays in advancing our understanding of ALS. Optimizing model architectures and integrating innovative techniques such as attention mechanisms enhance our ability to diagnose ALS and unravel the mechanisms underlying TDP-43 aggregation. These contributions are valuable for improving clinical outcomes and deepening our biological understanding of this devastating disease. As the field progresses, we anticipate that further collaborations between machine learning researchers, neuroscientists, and clinicians will drive significant advancements in diagnosing, monitoring, and treating ALS.

7 Conclusion and Future Work

In this study, we explored the performance of CNNs in the classification of ALS disease. We developed a model named Miniset-DenseSENet and evaluated its predictive performance using 190 autopsy brain images provided by the University of Aberdeen. Compared with standard classification models (such as ResNet18, DenseNet121, ResNet18+SE, and ResNet18+CBAM) and the baseline model (CNN), the Miniset-DenseSENet model significantly outperformed other models in classification accuracy on the small dataset.

Miniset-DenseSENet combines DenseNet121 with a SE module. The model's accuracy consistently exceeds 90%, with the best performance being 97.37%. The mean value of MCC is 0.84, and the optimal result is 0.96. The sensitivity of the control category is as high as 1, and the specificity at the worst performance is as high as 0.95. The main limitation of this model is its current computational complexity. However, this limitation can be addressed through methods such as model pruning, quantization, and knowledge distillation.

The development of the Miniset-DenseSENet model holds significant importance for ALS disease. Accurate classification and early diagnosis of ALS can lead to better patient management and treatment planning. With the use of advanced neural network techniques, this model has the potential to assist medical professionals in identifying ALS with high precision, thereby contributing to improved clinical outcomes and a deeper understanding of the disease.

Considering that the proposed method is trained on a limited dataset, although the model shows high accuracy, it still has deficiencies in terms of stability. To alleviate the overfitting problem and enhance the stability of the model, it is recommended that the size of the dataset be increased. Additionally, running the model multiple times and then evaluating the results using statistical tests is suggested to determine if the observed data indicate the presence of a true effect or if it may be a random event. While the model could not accurately distinguish the location of protein density in TDP-43 aggregates, the visualization images generated by Grad-CAM have been analyzed in collaboration with our clinical team. Further refinements in visualization techniques may be needed to improve the interpretability and diagnostic utility of the model.

A Additional Material

A.1 Image Quality Assessment

In this study, we evaluated the quality of the dataset using several key image quality metrics to ensure that the data is suitable for further analysis and deep learning training. The following sections describe the metrics used, along with the corresponding violin plots, which confirm that the images meet the required quality standards.

Laplacian Variance (Sharpness)

Laplacian variance is used to assess the sharpness of each image in the dataset. A high Laplacian variance value indicates sharper images, while a lower value suggests blurriness. Sharp images are crucial for identifying fine details in brain tissue samples, which are essential for accurate feature extraction in deep learning models.

The violin plot for Laplacian variance demonstrates that the majority of images have high sharpness, with only a few outliers in the lower range. This distribution confirms that the dataset predominantly consists of clear and sharp images that are suitable for analysis.

Brightness Average

Brightness average measures the overall intensity of each image, ensuring that the images are neither too dark nor too bright. Proper brightness is important for distinguishing between different regions and features in the images.

The brightness average violin plot shows that the images are well-distributed around a moderate brightness level, indicating that most images have appropriate lighting conditions. This ensures that brightness will not negatively affect the quality of data processing.

Contrast (Standard Deviation of Intensity)

Contrast, measured as the standard deviation of pixel intensity, reflects the difference between light and dark areas in each image. Adequate contrast is necessary to highlight important features and ensure that the images contain enough detail for deep learning models to extract relevant information.

The violin plot for contrast indicates a concentrated distribution of moderate to high contrast values, suggesting that the images exhibit sufficient variability between light and dark regions. This level of contrast is optimal for deep learning tasks.

Signal-to-Noise Ratio

SNR assesses the level of useful information in the image relative to background noise. High SNR values indicate clearer images with less noise, which is important for accurate classification and model performance.

The SNR violin plot shows that most images have a high SNR, with only a few low-SNR outliers. This confirms that the images are generally free from excessive noise, making them appropriate for deep learning training.

Based on the analysis of these quality metrics (Laplacian variance, brightness average, contrast, and SNR), the violin plots confirm that the dataset's image quality is acceptable. The majority of images exhibit sharpness, appropriate brightness and contrast, and low noise levels, ensuring that the dataset is valid for further data enhancement and deep learning model training.

B Confusion Matrices

References

[1] Arthur KC, Calvo A, Price TR, Geiger JT, Chio A, Traynor BJ. Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. Nature communications. 2016;7(1):12408.

[2] Tan RH, Ke YD, Ittner LM, Halliday GM. ALS/FTLD: experimental models and reality. Acta neuropathologica. 2017;133(2):177–196.

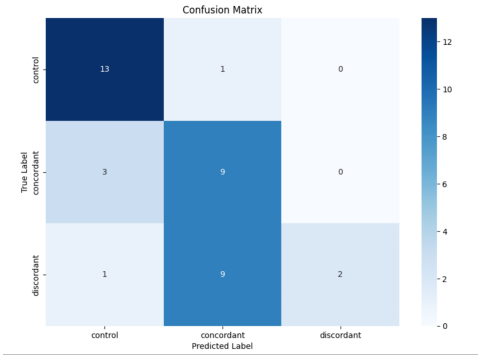
[3] Bede P. From qualitative radiological cues to machine learning: MRI-based diagnosis in neurodegeneration. Future Neurology. 2017;12(1):5–8.

[4] Gregory JM, McDade K, Bak TH, Pal S, Chandran S, Smith C, et al. Executive, language and fluency dysfunction are markers of localised TDP-43 cerebral pathology in non-demented ALS. Journal of Neurology, Neurosurgery & Psychiatry. 2020;91(2):149–157.

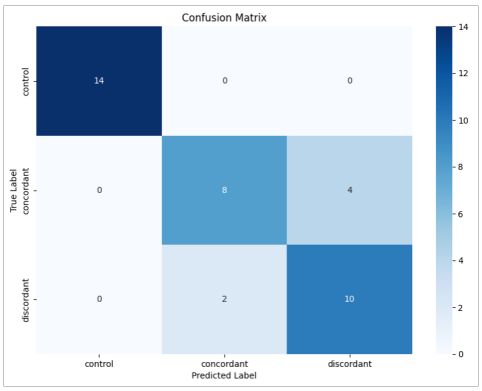
[5] Mejzini R, Flynn LL, Pitout IL, Fletcher S, Wilton SD, Akkari PA. ALS genetics, mechanisms, and therapeutics: where are we now? Frontiers in neuroscience. 2019;13:497022.

[6] JT C. Prognostic factors in motor neurone disease: A prospective study of longevity. Research progress in motor neuron disease. 1984;p. 34–43.

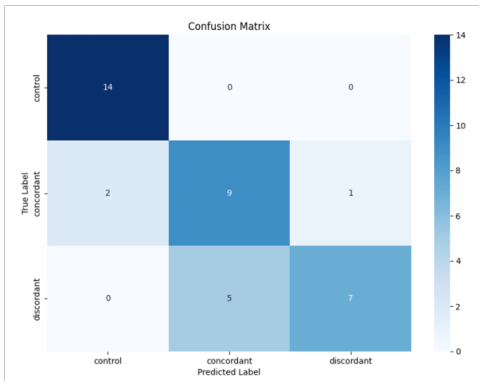
[7] Louwerse E, Visser C, Bossuyt P, Weverling G, et al. Amyotrophic lateral sclerosis: mortality risk during the course of the disease and prognostic factors. Journal of the neurological sciences. 1997;152:s10–s17.



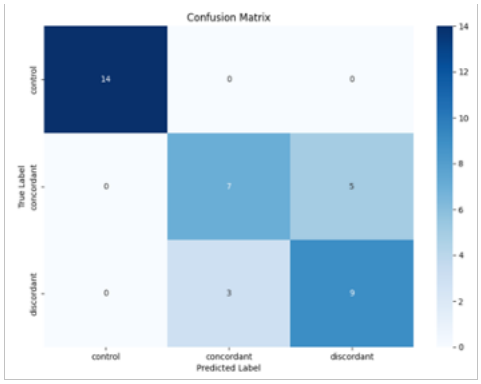
(a) Confusion matrices of basic CNN



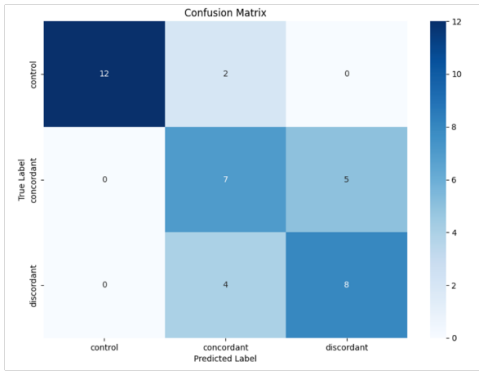
(b) Confusion matrices of ResNet18



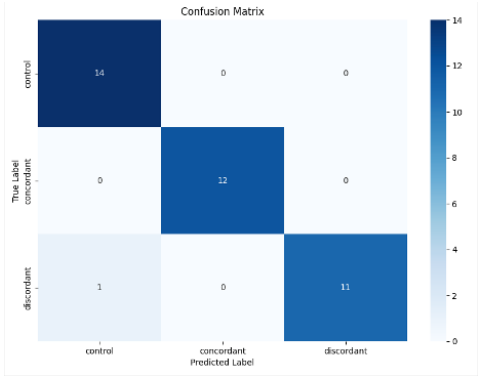
(c) Confusion matrices of DenseNet121



(d) Confusion matrices of ResNet18+SE



(e) Confusion matrices of ResNet18+CBAM



(f) Confusion matrices generated by the Miniset-DenseSENet.

Fig. 7: Confusion matrices of Different Models

[8] Hussain M, Bird JJ, Faria DR. A study on CNN transfer learning for image classification. In: Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK. Springer; 2019. p. 191–202.

[9] Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing. 2018;321:321–331.

[10] Grollemund V, Pradat PF, Querin G, Delbot F, Le Chat G, Pradat-Peyre JF, et al. Machine learning in amyotrophic lateral sclerosis: achievements, pitfalls, and future directions. Frontiers in neuroscience. 2019;13:438192.

[11] van der Burgh HK, Schmidt R, Westeneng HJ, de Reus MA, van den Berg LH, van den Heuvel MP. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. NeuroImage: Clinical. 2017;13:361–369.

[12] Liu X, Xing F, Yang C, Kuo CCJ, Babu S, El Fakhri G, et al. Voxelhop: Successive subspace learning for als disease classification using structural mri. IEEE journal of biomedical and health informatics. 2021;26(3):1128–1139.

[13] Thome J, Steinbach R, Grosskreutz J, Durstewitz D, Koppe G. Classification of amyotrophic lateral sclerosis by brain volume, connectivity, and network dynamics. Human brain mapping. 2022;43(2):681–699.

[14] Alzahrani AK, Alsheikhy AA, Shawly T, Azzahrani AS, AbuEid AI. Amyotrophic lateral sclerosis prediction framework using a multi-level encoders-decoders-based ensemble architecture technology. Journal of King Saud University-Computer and Information Sciences. 2024;p. 101960.

[15] Atmaramani R, Dreossi T, Ford K, Gan L, Mitchell J, Tu S, et al. Deep Learning Analysis on Images of iPSC-derived Motor Neurons Carrying fALS-genetics Reveals Disease-Relevant Phenotypes. bioRxiv. 2024;p. 2024–01.

[16] Han D, Liu Q, Fan W. A new image classification method using CNN transfer learning and web data augmentation. Expert Systems with Applications. 2018;95:43–56.

[17] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of big data. 2019;6(1):1–48.

[18] Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology. 2021;65(5):545–563.

[19] Huang Z, Du X, Chen L, Li Y, Liu M, Chou Y, et al. Convolutional neural network based on complex networks for brain tumor image classification with a

modified activation function. IEEE Access. 2020;8:89281–89290.

[20] Korolev S, Safiullin A, Belyaev M, Dodonova Y. Residual and plain convolutional neural networks for 3D brain MRI classification. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017). IEEE; 2017. p. 835–838.

[21] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big data. 2016;3:1–40.

[22] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems. 2019;32.

[23] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.

[25] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.

[26] Gangwar AK, Ravi V. Diabetic retinopathy detection using transfer learning and deep learning. In: Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1. Springer; 2021. p. 679–689.

[27] Khan S, Islam N, Jan Z, Din IU, Rodrigues JJC. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters. 2019;125:1–6.

[28] Deepak S, Ameer P. Brain tumor classification using deep CNN features via transfer learning. Computers in biology and medicine. 2019;111:103345.

[29] Chen L, Li S, Bai Q, Yang J, Jiang S, Miao Y. Review of image classification algorithms based on convolutional neural networks. Remote Sensing. 2021;13(22):4712.

[30] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48–62.

[31] Spence H, Waldron FM, Saleeb RS, Brown AL, Rifai OM, Gilodi M, et al. RNA aptamer reveals nuclear TDP-43 pathology is an early aggregation event that coincides with STMN-2 cryptic splicing and precedes clinical manifestation in ALS. Acta Neuropathologica. 2024;147(1):50.

[32] Zacco E, Kantelberg O, Milanetti E, Armaos A, Panei FP, Gregory J, et al. Probing TDP-43 condensation using an in silico designed aptamer. Nature Communications. 2022;13(1):3306.

[33] Waldron FM, Spence H, Gregory J. TDP-43 RNA aptamer staining to detect pathological TDP-43 in FFPE human tissue, as described in Spence and Waldron et al., 2024 (Acta Neuropathologica): A SOP and tick-sheet. v2. 2024;.

[34] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7132–7141.

[35] Amunts K, Zilles K. Architecture and organizational principles of Broca's region. Brain and Language. 2012;123(2):125–135.

[36] Fuster JM. The prefrontal cortex—an update: Time is of the essence. Neuron. 2001;30(2):319–333.

[37] Canny J. A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence. 1986;(6):679–698.

[38] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11):2278–2324.

[39] Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M. Medical image classification with convolutional neural network. In: 2014 13th international conference on control automation robotics & vision (ICARCV). IEEE; 2014. p. 844–848.

[40] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics. 2020;21:1–13.

[41] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 618–626.

[42] Wang Z, Bovik AC. Modern image quality assessment. Synthesis lectures on image, video, and multimedia processing. 2006;2(1):1–156.

[43] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019;6(1):1–54.

[44] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 2285–2294.