



Article

Autism Spectrum Disorder Detection Using a Weighted-Average Ensemble of Deep Convolutional Neural Networks on Eye-Tracking Images

Masroor Ahmed *, Sadam Hussain , Ivan Amaya and José Carlos Ortiz-Bayliss *

School of Engineering and Sciences, Tecnológico de Monterrey, Monterrey 64700, Mexico; a01753094@exatec.tec.mx (S.H.); iamaya2@tec.mx (I.A.)

* Correspondence: a00836487@tec.mx (M.A.); jcobayliss@tec.mx (J.C.O.-B.)

Abstract

Autism Spectrum Disorder is a long-term neurodevelopmental disorder. Early diagnosis is crucial for timely rehabilitation and intervention. Recently, machine learning and deep learning techniques have been widely explored and have produced encouraging results using eye-tracking scanpath images for the early detection of ASD. However, these approaches exhibit inconsistent performance and classification error rates, as well as limited generalization, due to differences in learning approaches and architectural designs across individual models. To address these problems, we employed a weighted-average ensemble of deep learning models using eye-tracking scanpath images. In this work, two different pretrained convolutional neural networks were selected, including Xception and VGG16, based on their proven efficacy and performance. Moreover, we fine-tuned each model individually and evaluated them on a dataset containing eye-tracking scanpath images. We implemented a weighted-average ensemble technique to combine the diverse predictions of the two models. It reduces classification errors and improves the model's generalization and overall performance. The adopted weighted-average ensemble technique achieved an accuracy of 98.18%, with a perfect recall, and a competitive Area Under the Curve (AUC) of 99.59%. These findings highlight that applying a weighted average to integrate multiple models' predictions strengthens the generalization and reliability of ASD detection.

Keywords: ensemble learning; deep learning; weighted average; Autism Spectrum Disorder (ASD)

1. Introduction

Developmental disorders are long-term impairments that have a significant influence on the day-to-day functioning of a great number of people. Among those developmental disorders, Autism Spectrum Disorder (ASD) is one of the most heterogeneous. The number of individuals living with ASD is rising at an alarming rate worldwide [1,2].

The term 'autism' was initially used to describe a behavioral manifestation of self-withdrawal observed in individuals with schizophrenia. In 1943, Kanner and Asperger coined the term 'autism' to characterize a syndrome observed in children who exhibited distinct behavioral variations in social interaction and communication and a tendency towards restricted and repetitive interests [3,4]. ASD is a term that refers to a complex neurodevelopmental illness characterized by continual or chronic deficiencies in verbal and nonverbal, social communication, social interactions, and limited, repetitive patterns of behavior, hobbies, and activities [5,6].



Academic Editor: Angelo Cangelosi

Received: 28 May 2026

Revised: 18 June 2026

Accepted: 23 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

In contemporary times, the inclusive classification ASD categorizes a diverse range of neurodevelopmental disorders characterized by shared fundamental behavioral traits impacting social communication, as well as the presence of restrictive and repetitive stereotypical behaviors and interests. The term ASD encompasses various illnesses, including idiopathic forms of autism, Asperger syndrome, Pervasive Developmental Disorder-Not Otherwise Specified, and Childhood Disintegrative Disorder. Additionally, certain hereditary disorders, such as Rett syndrome, may manifest autistic features [5,7]. In the past few decades, the number of people with ASD has grown worldwide. Based on the data published by the World Health Organization (WHO), it is estimated that around 1% of children all over the world are affected with ASD. Moreover, according to the Centers for Disease Control and Prevention (CDC), in 2023, the prevalence of ASD was estimated to be one of the highest rates in the United States, with around 1 in 36 children being affected [8].

We commonly attribute the etiology of ASD to a combination of genetic and environmental influences. Although it may be possible to identify the features of a child with ASD, they are more commonly observable in early childhood. However, the diagnosis of ASD often remains undetermined until a later stage. The implementation of timely identification and interventions can effectively enhance the overall physical and mental health, as well as the overall quality of life, for those affected by ASD. Additionally, these interventions have the potential to positively impact the cognitive development of individuals with ASD [9,10].

The Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview (ADI) are examples of conventional techniques for diagnosing ASD. These techniques entail behavioral and observational assessments carried out by parents and doctors. These conventional techniques mostly depend on behavioral evaluations and observational tools, which can be time-consuming and inherently subjective [8,9,11]. However, many researchers are working on studies that focus on establishing interactive tools such as Magnetic Resonance Imaging (MRI), electroencephalography (EEG), facial expression, genetic testing, etc., to promote the treatment and rehabilitation of people with ASD, as well as automated computer-aided methods for detecting the condition, considering current technology breakthroughs. These automated techniques are essential for promoting early detection, lowering subjectivity, and improving the reliability and accessibility of diagnosis [11].

Eye tracking has received increasing attention in recent years for capturing objective behavioral biomarkers of visual attention in ASD and represents a useful technique for analyzing ASD. It is a technology that can record, monitor, and quantify eye movements or absolute points of gaze (POG), which is the point where the eye gaze focuses on a visual scene [12]. Eye tracking aims to identify abnormal visual attention that contributes to the clinical characteristics of ASD and to analyze its perceptual components. Eye tracking patterns in people with ASD significantly differ from those of non-ASD individuals. Therefore, the current literature recommends incorporating eye-tracking scanpaths (ETSP) data into the detection and diagnosis of ASD [13,14].

Recent developments in the application of Artificial Intelligence (AI) in healthcare have produced promising results. It is becoming common to rely on AI-driven methods to automate complex diagnostic processes and extract meaningful patterns from high-dimensional feature spaces, thereby providing vital decision support to clinicians [15]. The literature provides examples of several Machine Learning (ML) based computational models that successfully identify patterns in health-related clinical and behavioral data [16–18]. Moreover, prior research shows that ML pipelines can improve ASD detection by learning subtle behaviors, varied symptoms, and previously unexplored cues from diverse patient data [19,20]. Additionally, with the introduction of Deep Learning (DL) architectures, peo-

ple have developed more robust algorithms for healthcare in general and ASD detection in particular. For example, by using Convolutional Neural Networks (CNNs), we are capable of extracting spatial local features through their learnable filter in contrast to the manual feature engineering of linear ML models [21].

Overall, researchers have relied on ML and DL approaches to automate the detection of ASD using eye-tracking scanpath data. Specifically, deep learning architectures, such as Deep Convolutional Neural Networks (DCNNs), have been evaluated using these inputs to automatically extract discriminative features, demonstrating encouraging performance [13,14,22]. However, there are inherent drawbacks to using DL models, especially when the training data is small. Among the drawbacks, we can mention overfitting, architectural bias, and limited feature diversity. As a result, such drawbacks can limit the model's generalizability and reliability when a single architecture is used [23,24]. To address these challenges, Ensemble Learning (EL) has emerged as a viable alternative that effectively leverages knowledge from multiple models [24]. Ensemble Learning (EL) combines the individual strengths of models. It compensates for their weaknesses, reducing reliance on a single, potentially biased model and enhancing both generalization and predictive accuracy relative to individual models [25–27]. In this study, we employ a weighted-average ensemble technique to address these problems by integrating the complementary strengths of two models, Xception and VGG16, to improve detection performance on eye-tracking scanpath images and enhance the model's generalization and robustness.

The main contributions of this work are as follows:

- We propose a DL-based weighted-average ensemble framework for ASD detection that combines the prediction probabilities of Xception and VGG16 using validation-optimized weights, providing a transparent and computationally efficient alternative to more complex ensemble schemes.
- We compare the findings of the proposed framework against those of state-of-the-art methods and individual DCNN models with and without augmentation.
- We present class-wise performance for individual models and the ensemble, yielding detailed insights into class-specific behavior and highlighting diagnostic reliability.

We have organized the remainder of this document as follows. Section 2 presents the work related to this investigation. Section 3 discusses the proposed deep learning-based weighted average ensemble technique for ASD detection, including the dataset, preprocessing techniques, classifiers, ensemble strategy, and evaluation metrics. Section 4 presents the results and discussion. Finally, Section 5 presents the conclusion and some ideas for future work.

2. Related Work

Several studies have investigated how ML and DL models can be used to detect ASD, leveraging various techniques and algorithms to improve performance. The following subsections summarize the most relevant works on ML and DL applied to ASD detection or identification using eye-tracking data.

2.1. The Visualization of Eye-Tracking Scanpaths Image Dataset

Before reviewing the related studies, we describe the dataset used in most of them. Moreover, we will also consider this dataset in the experiments conducted in this work. This dataset, known as the Visualization of Eye-Tracking Scanpaths in Autism Spectrum Disorder Image Dataset [28,29], is publicly available from the Figshare data repository (<https://doi.org/10.6084/m9.figshare.7073087>). Their authors, Carrette et al. [28,29], collected data from 59 children aged 3–13 years (38 males and 21 females), of whom 30 had ASD, and 29 were typically developing (TD). The dataset contains 547 images: 328 classified

as TD and 219 as ASD. This dataset supports binary classification because it only contains two classes: ASD and TD. From this point on, we refer to these data as the VETSP image dataset. Also, we provide some sample raw images (Figure 1), for the sake of clarity.

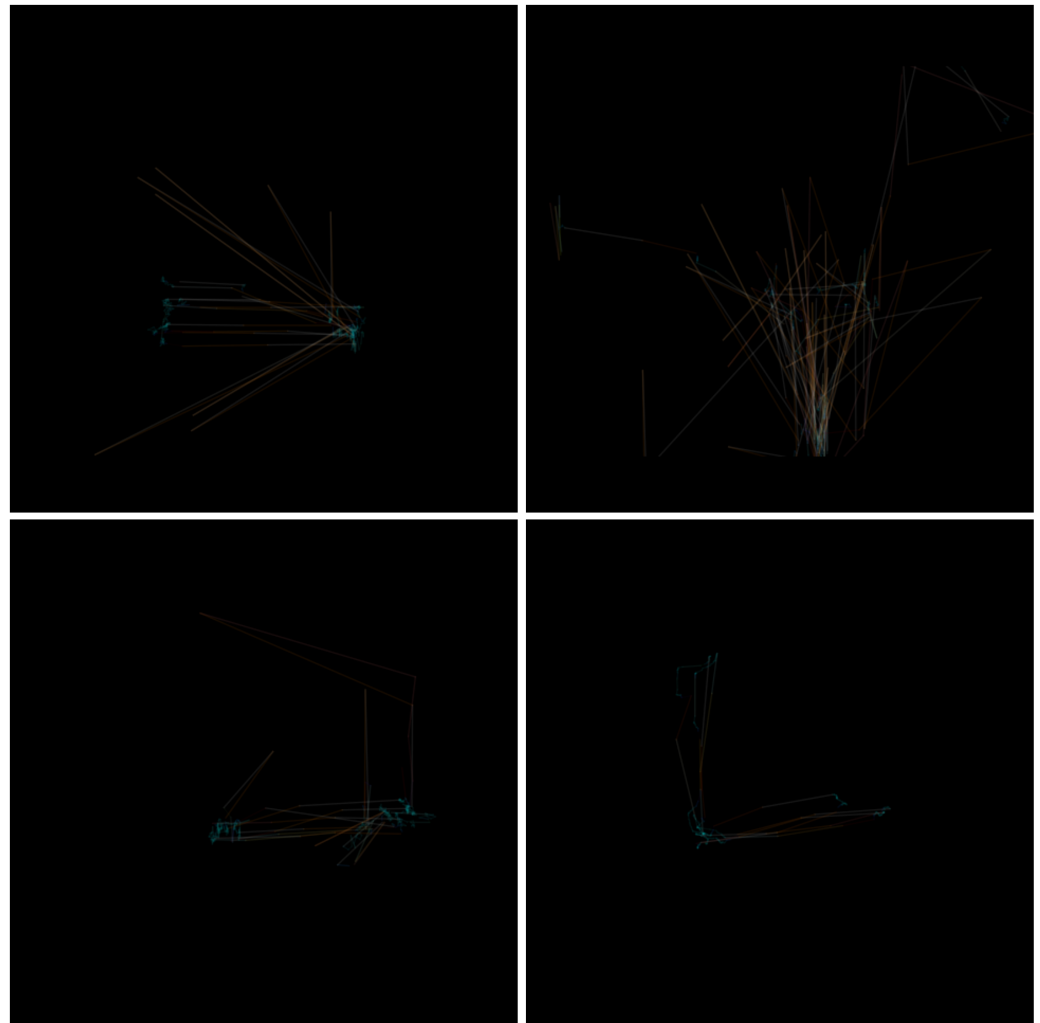


Figure 1. A few sample images from the VETSP image dataset.

2.2. Traditional Machine Learning Approaches on Eye-Tracking Data

Early research in ASD detection focused primarily on conventional supervised ML techniques applied to eye-tracking data, emphasizing feature engineering, dimensionality reduction, and classical classifiers. For example, Carette et al. [29] proposed a method that uses ML classifiers such as Logistic Regression (LR), Naïve Bayes Classifier (NB), Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN) to identify ASD in children by extracting eye-tracking patterns that correlate with ASD symptoms. Carette et al. [29] used preprocessing techniques, data augmentation, and Principal Component Analysis (PCA) on the VETSP image dataset. Among these models, the ANN achieved the best performance, with 92.00% accuracy and an AUC of $92.00\% \pm 0.02$ for binary classification.

A few years later, Akter et al. [30] introduced an ML model for early detection of ASD using the VETSP image dataset [29]. To address high image similarity, the authors used *k*-means clustering to split their data into four clusters, thereby improving classifier performance. Akter et al. [30] trained multiple ML classifiers, including Decision Trees (DT), Gradient Boost (GB), *k*-Nearest Neighbors (KNN), LR, Multilayer Perceptron (MLP), NB, RF, SVM, and Extreme Gradient Boosting (XGB). The authors evaluated these models on

both their primary dataset and clustered subsets. Among all models, MLP achieved the highest performance at the cluster level, but RF outperformed the others on the complete dataset, with 74.22% accuracy and 71.53% AUC.

More recently, Kanhirakadavath and Chandran [13] used ML techniques to identify ASD in children, using the VETSP image dataset [29], and employed two separate frameworks. The first framework included data preprocessing, PCA for dimensionality reduction and feature extraction, and several ML classifiers: RF, DT, SVM, LR, Decision Jungle (DJ), Deep SVM, and Boosted Decision Tree (BDT). In contrast, the second framework used a DL strategy, training a Deep Neural Network (DNN) and using CNNs for feature extraction. Notably, the CNN-based DNN achieved 97.00% AUC and outperformed all classical models on the test cases.

Although most studies focus on visual attention patterns, some researchers have explored alternative modalities, such as speech and acoustic features, to complement visual-based diagnosis. Li et al. [31] presented a novel automated system that measures stereotyped and idiosyncratic speech patterns, as well as abnormal prosody, in patients with ASD. The authors implemented two approaches simultaneously, including end-to-end DL methods and manually designed features. The researchers started using the OpenSMILE toolbox to extract high-dimensional acoustic properties from utterances, then implemented an SVM as their standard baseline. The authors modeled unusual prosody labels from voice spectrograms using DNNs with designated design parameters. The DL approach, when combined with the traditional baseline at the score level, improved the system's performance to 85.70% accuracy for stereotyped and 88.10% accuracy for atypical prosody in ASD binary classification. They collected a speech (audio) dataset and utilized it to analyze atypical prosody and stereotyped phrases in children with ASD during Autism Diagnostic Observation Schedule (ADOS) screening sessions at the affiliated hospital of Sun Yat-sen University.

2.3. Unsupervised and Representation Learning Approaches

In parallel with supervised learning methods, some studies have explored unsupervised representation learning techniques to uncover latent behavioral patterns without relying on labeled data. Among those works, it is worth mentioning the one by Elbattah et al. [32], who presented an unsupervised ML approach to analyze ASD on the VETSP image dataset [29], extracting features via a deep autoencoder. Building on these latent representations, Elbattah et al. [32] clustered them with k -Means to uncover gaze-behavior patterns without supervised labeling. Overall, their study presented deep autoencoders with scanpath visualization, enabling the unsupervised discovery of ASD-related patterns in eye-tracking data and supporting the development of future detection tools.

2.4. Convolutional and Deep Neural Network-Based Models

With the growing availability of computational resources and larger datasets, research progressively shifted toward DL architectures, particularly convolutional and hybrid neural networks. An early example of these approaches is the work by Cilia et al. [33], who proposed a CNN model for ASD screening on the VETSP image dataset [29]. Their model comprises four convolutional layers, four pooling layers, and two fully connected layers for image classification, with dropout layers implemented to reduce overfitting. In their work, the model achieved an accuracy of 90.00% and an AUC of $90.00\% \pm 0.01$, while participant-wise analysis yielded an accuracy and an AUC of $71.00\% \pm 0.01$. Furthermore, a maximal information coefficient (MIC) study found a strong nonlinear relationship ($\text{MIC} = 0.79$) between ASD severity and eye movement speed. These findings illustrate that CNNs are

reliable for identifying unusual attention patterns in people with ASD, suggesting that such methods could serve as potential non-invasive screening tools.

A few years later, Kabir Mehedi et al. [34] proposed a classification method for ASD based on CNNs and a Vision Transformer architecture, trained and validated on two eye-tracking scanpath datasets: VETSP [29] and VAF-ASD [35]. Kabir Mehedi et al. [34] evaluated several CNN models, including InceptionV3, EfficientNetB7, MobileNetV2, and Xception. Besides, they employed a lightweight Vision Transformer model, the Compact Convolutional Transformer (CCT), with 0.4 million parameters. The proposed CCT model achieved accuracies of 97.62% on the VETSP and 96.43% on VAF-ASD, outperforming the other evaluated models.

More recently, Alsaidi et al. [14] proposed a DL model called T-CNN-ASD for ASD detection using the VETSP image dataset [29]. The T-CNN-ASD architecture comprises multiple layers, including convolutional layers, ReLU activations, pooling, and fully connected layers. After flattening, the two fully connected hidden layers with 300 and 150 neurons, respectively, were used. However, to achieve optimal model performance, they fine-tuned the hyperparameters. The model achieved 95.59% accuracy in 10-fold cross-validation and was compared with ML models (DT, RF, KNN, and MLP) without data augmentation. The results highlight the model's ability to identify nuanced eye-tracking patterns. This indicates significant potential to improve early clinical screening for ASD and support timely intervention.

Almost simultaneously, Mumenin et al. [36] presented a lightweight Involutorial Neural Network model for ASD detection. It required fewer parameters than state-of-the-art image classification models. The proposed model used involution kernels rather than standard convolutions, enabling it to recognize location-specific spatial patterns successfully. As a result, ASDNet achieved an accuracy of 98.12% and an AUC of 99.72% on the VETSP image dataset [29], while in VAF-ASD, it achieved an accuracy of 96.83% and an AUC of 99.10% for fixation maps and an accuracy of 97.61% and an AUC of 96.29% for heatmaps. These results surpassed CNN and transformer baselines. Additionally, Monte Carlo Dropout was applied to the model, enabling the calculation of prediction uncertainty and confirming model reliability.

Finally, Soloh et al. [37] proposed a lightweight DL method for early ASD detection. They used the VETSP image dataset [29], and data augmentation to improve the model's reliability and generalization. They also applied five-fold stratified cross-validation and provided a robust benchmark for gaze-based detection. Moreover, a 16-point geometric calibration protocol was applied to standardize gaze vectors. This enhanced spatial reliability across diverse recording conditions. The proposed model achieved an accuracy of 97.01% and an AUC of 98.00%. It also supports real-time mobile deployment and demonstrates the framework's feasibility as a low-cost, accessible solution for early ASD screening.

2.5. Ensemble and Hybrid Learning Frameworks

Jaradat et al. [22] developed a hybrid ensemble model using the VETSP image dataset [29]. They applied preprocessing, data augmentation, extracted features using MobileNet, and applied PCA to reduce dimensionality. For classification, they constructed a stacking ensemble in which predictions from classifiers, SVM, KNN, and deep learning models (MobileNet, VGG19, Xception, InceptionV3) were combined by a meta-learner to make the final decision. The hybrid model with MobileNet achieved accuracies of 96.17% for eye gaze maps, 98.02% for ETSDs, and 98.18% for ETSP images. These findings demonstrate that integrating deep feature extraction within an ensemble learning approach yields the most consistent diagnostic performance across datasets.

In their study, Ahmed et al. [38] introduced an ensemble model based on the VETSP image dataset [29] using DL for ASD detection. They improved visual quality and feature extraction using preprocessing techniques such as histogram equalization and enhanced model generalization with data augmentation. Pretrained models, including DenseNet169, VGG19, and MobileNet, achieved an accuracy of 95.00% to 96.00%. Their ensemble approach concatenated features from VGG19, DenseNet169, MobileNet, and ViT, reaching 98.00% accuracy for binary classification, outperforming individual models and demonstrating the effectiveness of model concatenation.

2.6. Few-Shot and Contrastive Learning Approaches

More recent work has begun to address data scarcity and generalization limitations through few-shot and contrastive learning paradigms. Mousli et al. [39] proposed ConASD, a contrastive few-shot learning approach for detecting ASD using two real-world eye-tracking image datasets. The authors employed a two-stage process: in the first stage, the ResNet50 encoder is trained to extract discriminative and compact features via supervised contrastive learning, thereby improving inter-class separation. In the second stage, the encoder is fine-tuned using cross-entropy loss for ASD classification. When they trained the proposed model on the full dataset, it achieved accuracies of 69.73% and 76.67% on the VETSP image dataset [29] and on the VAF-ASD dataset [40], respectively. The findings highlight that the proposed model outperformed ML models and standard CNN training models.

3. Proposed Framework

This section describes the elements of the proposed framework for ASD detection based on the VETSP image dataset. This study uses a weighted-average ensemble technique that combines predictions from two DCNN architectures, Xception and VGG16, based on the validation dataset. We selected these architectures due to their complementary structural designs, proven efficacy, and performance. Each model is separately trained and fine-tuned on this dataset to ensure unbiased feature representation and standardized assessment. After training and fine-tuning, we integrated the probabilistic classification outputs using a weighted-average ensemble strategy, with optimal weights determined on a validation set to obtain final predictions based on grid search. This framework leverages the complementary features of both architectures, yielding overall performance, improved model generalizability, reduced overfitting risk, and enhanced robustness relative to the individual models. Figure 2 illustrates the proposed framework.

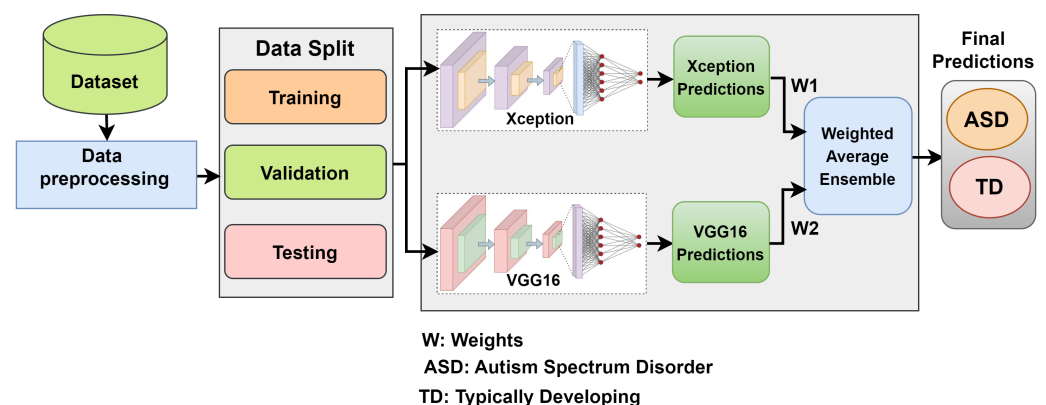


Figure 2. The proposed framework includes image preprocessing, individual model training, evaluation, and combining the prediction probabilities of both models using a weighted average to obtain the final predictions.

3.1. Data Preprocessing and Dataset Split

Before feeding the deep learning model the training images, it is critical to preprocess the VETSP image dataset. To this end, this work employed various preprocessing techniques on the ASD scanpath data to enhance its quality and stability, thereby improving feature learning and model convergence. Initially, all images were resized to 128×128 pixels to standardize spatial dimensions, minimize computational complexity, and ensure uniform input sizes across the employed DCNN architectures. Afterwards, we applied various image enhancement techniques, such as contrast and sharpness adjustment, to preserve structural boundaries. To compensate for brightness inconsistencies and illumination variations, we relied on gamma correction to improve under-exposed regions and a power-law transformation to correct over-exposure. After gamma correction, we converted the images to grayscale and applied morphological operations to refine structural patterns and minimize small-scale noise irregularities. We also used these operations to thicken prominent edges, fill small gaps, strengthen weak areas, and reduce small inconsistencies. Morphological operations produce the necessary stability and discriminative power in the data for better model training. Furthermore, we normalized pixel intensities to the common range $[0, 1]$ to ensure network stability and accelerate convergence. We replicated the grayscale images into three channels to create a pseudo-RGB input without introducing new color information.

After preprocessing, the dataset was split into training, validation, and test sets with proportions of 80%, 10%, and 10%, respectively. A similar dataset partitioning strategy has been utilized in a previous deep learning study to facilitate effective learning, parameter tuning, and consistent model evaluation [36]. Because the VETSP dataset contains a limited number of participants but a larger number of scanpath images, we adopted an image-level split rather than a participant-level split. A participant-level partition would have produced relatively small validation and test subsets, potentially affecting class balance and the stability of model comparison. For this reason, we used stratified random sampling with a fixed random seed of 32 to preserve the class distribution across the training, validation, and testing subsets and to ensure reproducibility. In addition, we stored the image paths for each subset in separate CSV files and reused them consistently across all evaluated models, enabling a fair comparison under the same data partition. Finally, to mitigate overfitting and improve the model's generalization, we applied simple data augmentation transformations to the training dataset: rotations, width shifts, height shifts, and horizontal flips. By applying such preprocessing techniques, we improved the overall quality of the data, enabling more consistent and more effective learning. We present all preprocessing parameters in Table 1.

Table 1. Summary of the preprocessing and dataset split parameters.

Parameter	Values
Resize image	128×128
Image enhancement	Contrast = 2.0, Sharpness = 6.2
Gamma value	2.2
Morphological operations	kernel 3×3 ; dilation 2 iteration followed by closing
Normalization	Min–max normalization in the range $[0, 1]$
Color format	Grayscale (replicated to three channels)
Seed	32
Data split ratio	80% (train)/10% (validation)/10% (test)
Augmentation (training only)	Rotation (20), width/height shift (0.2), horizontal flip

3.2. Deep Convolutional Neural Networks

Deep learning (DL)-based models, particularly CNNs, have shown remarkable efficiency in tackling image classification problems. Although CNNs are most commonly utilized for visual data, their robust capability to learn rich feature representations also makes them applicable to a wide range of data analysis tasks. In this study, we used two state-of-the-art DCNN architectures, Xception and VGG16, to detect ASD using the VETSP image dataset and evaluated their performance individually. We provide additional descriptions of these models below.

The Xception model. Google researchers introduced Xception, a DCNN architecture known as Extreme Inception [41]. Xception decouples spatial and cross-channel feature learning by replacing Inception modules with depthwise separable convolutions. It relies on residual connections to enhance convergence and consists of a linear stack of depthwise separable convolution layers arranged into entry, middle, and exit flows. Xception has a similar number of parameters to InceptionV3 but performs better by using its parameters more efficiently. In our work, we use the Xception model as the base architecture for feature extraction on the VETSP image dataset and utilize pretrained ImageNet weights to obtain robust, generalized features and reduce the need for training from scratch. We flattened the extracted features and passed them to a classification head with two fully connected layers of 512 and 256 neurons, each using the ReLU activation function. Thus, we modified the top layers of the Xception model to include additional fully connected layers, particularly for the intended dataset. We also used dropout layers with rates of 0.5 and 0.3 to reduce overfitting and boost generalization. The final layer consisted of two neurons with softmax activation. For training, we relied on the Adam optimizer with a learning rate of 0.001. We used categorical cross-entropy as the loss function during model training and trained for 100 epochs, using a batch size of 32 to ensure reliable convergence.

The VGG16 model. VGG16 is a DCNN architecture introduced by Simonyan and Zisserman [42]. It was trained and tested on the ImageNet dataset, which comprises 1000 classes and more than a million images. VGG16 is composed of 16 weighted layers: 13 convolutional layers with 3×3 filters and three fully connected (dense) layers. The network also includes five max pooling layers that follow certain convolutional layers. In the fully connected part, the first and second layers each contain 4096 neurons, the third layer contains 1000 neurons, and the output is a softmax activation layer.

This work uses the VGG16 model, pretrained on the ImageNet dataset, for feature extraction from the VETSP image dataset for ASD classification. We modified the top layers of VGG16 to include additional fully connected classification layers tailored to the targeted dataset. We passed the extracted features to fully connected layers with 1024, 512, and 256 neurons, each with a ReLU activation function to enhance the model's nonlinearity and facilitate learning. Dropout regularization with rates of 0.4 and 0.5 was applied to the dense layers to alleviate overfitting. We used the Softmax function in the classification head to produce final class probabilities, with two values indicating either ASD or TD. The model used the Adam optimizer with a learning rate of 0.001 and the categorical cross-entropy loss function during training. The training was carried out for 100 epochs using a batch size of 32.

For clarity, Table 2 presents the parameters used for these CNN models in this work.

Table 2. Model architecture and training configuration of the evaluated CNN models.

Parameter	Xception	VGG16
Base model	Xception	VGG16
Pretrained weights	ImageNet	ImageNet
Input size	$128 \times 128 \times 3$	$128 \times 128 \times 3$
Frozen layers	All backbone layers	All backbone layers
Fully connected layers	512, 256	1024, 512, 256
Activation function	ReLU	ReLU
Output layer	2 neurons, Softmax	2 neurons, Softmax
Dropout rates	0.5, 0.3	0.4, 0.5
Optimizer	Adam	Adam
Learning rate	0.001	0.001
Loss function	Categorical cross-entropy	Categorical cross-entropy
Batch size	32	32
Epochs	100	100

3.3. Ensemble Learning

Ensemble learning (EL) is an effective method for image classification because it integrates predictions from multiple models rather than relying on a single model. The basic concept behind EL is that each model interprets the data in a slightly different way and possesses its own strengths and weaknesses [43]. When these models act together, the combined predictions are more accurate, stable, and reliable than those of any individual model. Additionally, EL helps prevent common issues such as overfitting, noise-related errors, and biased predictions, thereby enhancing performance on difficult or noisy datasets, such as medical or real-world images [44]. Among the widely employed ensemble techniques, we find bagging, boosting, and stacking, each of which improves performance in different ways. Bagging creates multiple training data subsets and trains a model on each. Combining their outputs assists in reducing overfitting and prevents all models from repeating the same mistake. In boosting, the models are trained one after another, and each model attempts to address the errors produced by earlier models [43,45]. As more models focus on difficult samples, the ensemble becomes more accurate. While bagging reduces variance, boosting reduces bias by learning from past errors. Stacking involves training multiple models and combining their knowledge into a new model. The resulting combined model combines the best learning paradigms of the individual models and is robust enough to yield improved performance on evaluation metrics. To this end, EL has the potential to improve classification performance by combining the weights learned by individual models, thus fusing the learning mechanisms of diverse feature extraction algorithms. The field of EL has proven to be an important research domain in computer vision and medical imaging, as it can reduce model errors and alleviate overfitting. Although various ensemble approaches, such as bagging [46], boosting [47], and stacking [48,49], are popular in the literature, simpler techniques, such as averaging and weighted averaging [50,51], have also proved effective in image classification tasks, offering improved performance while being easy to implement.

The averaging ensemble method aims to combine predictions from multiple models by averaging their outputs. This implies that the individual model probabilities are combined and divided by the total number of models participating in the ensemble. We computed the final class predictions as the highest averaged class probability. We can represent the average ensemble as given in Equation (1).

$$y^* = \arg \max_i \frac{1}{n} \sum_{j=1}^m w_{ij} \quad (1)$$

where w_{ij} denotes the probability determined for the i -th class label by the j -th classifier [52].

Weighted average voting is an extended version of the simple average method, where each model is assigned a different weight based on its importance or performance, as shown in Figure 3. Instead of treating all models equally, each prediction is multiplied by its corresponding weight to form a weighted sum. This sum is then divided by the total of the weights to obtain the final weighted average for each class, as displayed in the function. This technique allows individual models to contribute more to the final decision, often achieving higher accuracy than the simple average method [52]. The weighted average formula is displayed in Equation (2).

$$y^* = \frac{\sum_{j=1}^m w_j x_j}{\sum_{j=1}^m w_j} \quad (2)$$

where y^* denotes the weighted average, m shows the number of terms to be averaged, the weights applied to x values are w_j , and the data values to be averaged are x_j [52].

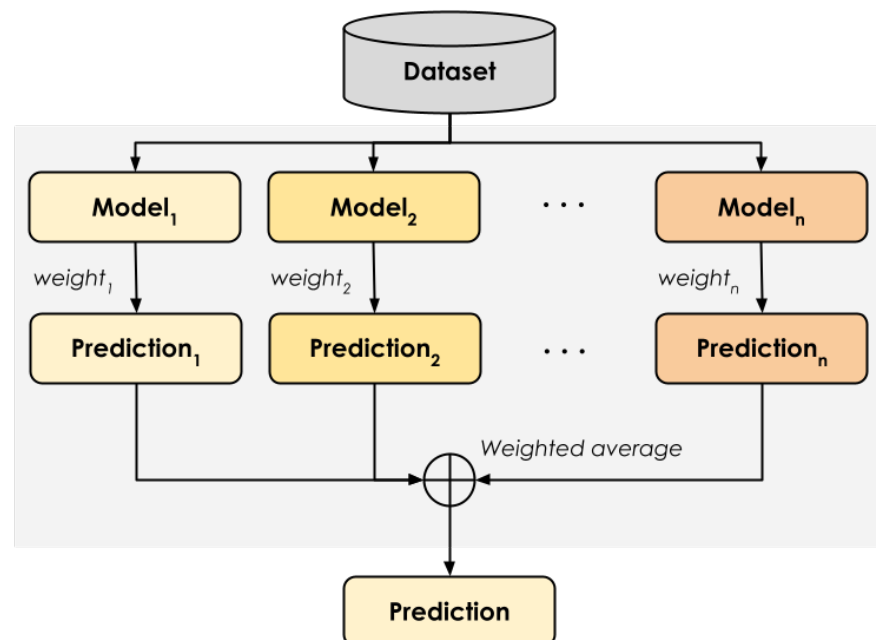


Figure 3. The figure represents the weighted-average ensemble technique, where prediction outputs from multiple classifiers are integrated based on assigned weights to achieve the final output prediction.

The novelty of the adopted WA-Ensemble approach lies in its ability to integrate the prediction probabilities of multiple models while addressing several limitations of existing ensemble techniques. For example, in the straightforward average technique, equal weights are assigned to all base models regardless of their individual predictive performance, potentially diminishing the influence of stronger classifiers [53]. Bagging enhances robustness but has limitations, including increased computational and memory requirements and an inability to account for differences in model outcomes explicitly [52,54]. Although boosting tends to improve weak learners, it increases training complexity and the risk of overfitting, especially with limited datasets [52,53]. Similarly, the stacking ensemble approach relies on an additional meta-classifier and a supplementary training stage, thereby increasing computational cost and limiting interpretability [53,54].

With our WA-Ensemble technique, prediction probabilities are combined using validation-optimized weights, allowing the model with stronger validation performance

to contribute more effectively to the final prediction. Moreover, it is a simple and efficient approach with lower computational cost than its counterparts, as it does not involve additional processing stages and improves reproducibility. It also provides greater transparency, making it particularly suitable for clinical decision support applications [53].

Overall, the novelty of the WA-Ensemble lies in its transparent probability-driven weighting strategy, which allows stronger classifiers to exert greater influence on the final decision without requiring additional meta-learning or resampling stages. By combining predictive performance, computational simplicity, and interpretability, the method offers an efficient and clinically suitable alternative to conventional ensemble schemes [53,54].

In this study, we used the weighted-average ensemble technique to improve the model's overall performance for ASD classification on the VETSP image dataset. Two pretrained DCNN models, Xception and VGG16, were first trained and evaluated individually to produce prediction probabilities. To integrate the models, we combined their output probabilities using a weighted average: we multiplied each model's probability by its assigned weight and summed the results. This integration leverages the complementary strengths of the individual models, aiming to reduce classification error rates and enhance reliability, generalization, and overall performance. The weights were determined via a grid search over 0 to 1 in increments of 0.01, with the constraint that the total weight equals 1. Optimal weights, selected based on validation set performance, were then applied to combine the models' predictions for final classification.

3.4. Performance Evaluation Metrics

In this section, we describe the performance metrics considered for this work: accuracy (Equation (3)), precision (Equation (4)), recall (Equation (5)), and F1-score (Equation (6)).

$$\text{Accuracy}(\%) = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Precision}(\%) = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall}(\%) = \frac{TP}{TP + FN} \quad (5)$$

$$\text{F1-score}(\%) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

where true positives (TP) refer to the number of positive instances correctly identified as positive, false negatives (FN) indicate the number of positive instances incorrectly categorized as negative, true negatives (TN) correspond to the number of negative instances correctly identified as negative, and false positives (FP) represent the number of negative instances incorrectly classified as positive.

4. Experiments and Results

4.1. Baseline Performance of Fine-Tuned Xception and VGG16 Models

Initially, we used some preprocessing steps to improve the image quality and ensure consistency. Then, we fine-tuned two pre-trained classifiers, namely Xception and VGG16, and used them as baselines. We combined the predictions from these two models using a weighted average, with weights optimized on the validation dataset. We evaluated the performance of these models using accuracy, precision, recall, and F1-score. The Xception model achieved 96.36% accuracy, with balanced results of 95.45% precision, recall, and F1-score. Conversely, the VGG16 model also achieved 96.36% accuracy and reported lower precision (91.67%) and perfect recall (100.00%), demonstrating a slight improvement in

F1-score (95.65%). Although both models performed competitively and robustly overall, their complementary qualities encourage their integration in an ensemble framework.

Aiming for a deeper analysis, we present the confusion matrices for the Xception and VGG16 models in Figure 4. The Xception model shows a balanced classification performance, with similar error rates for both classes. It accurately classified 32 TD samples, inaccurately categorizing only one TD case as ASD. In the same way, it correctly classified 21 ASD samples, while mistaking one ASD case as TD. The model's low false-positive and false-negative rates indicate a stable trade-off between precision and recall, suggesting it can consistently distinguish between ASD and TD classes. The VGG16 model, on the other hand, correctly identified 31 TD cases and misclassified two as ASD. All 22 ASD samples were correctly classified, obtaining no false negatives. These findings suggest that the model is competent at detecting ASD, since it did not miss any positive cases. Still, the fact that there are a few more false positives than Xception implies that there is a small propensity to over-classify ASD cases.

		Xception				VGG16	
		TD	ASD			TD	ASD
Actual class	TD	32 58.18%	1 1.81%	Actual class	TD	31 56.36%	2 3.64%
	ASD	1 1.81%	21 38.18%		ASD	0 0.00%	22 40.00%
		Predicted class				Predicted class	

Figure 4. Confusion matrices of the baseline models (Xception and VGG16) for ASD detection on the test set.

4.2. Performance Evaluation of a Weighted-Average Ensemble of Xception and VGG16

To leverage the diverse predictive strengths of two DCNN architectures, we employed a weighted-average ensemble to integrate the complementary strengths of Xception and VGG16, thereby enhancing model performance and generalization. Using the weighted-average ensemble technique, we assigned optimized weights to the individual models based on the validation dataset to fine-tune the ensemble and improve its performance. The optimized weights are 0.37 for Xception and 0.63 for VGG16, indicating their relative contributions to the final performance. The efficacy of the weighted-average ensemble approach—hereafter referred to as WA-Ensemble—is demonstrated compared to the individual models Xception and VGG16. As depicted in Table 3, the individual models Xception and VGG16 achieved the same accuracy of 96.36%. Xception had balanced precision and recall of 95.45%, while VGG16 had perfect recall of 100.00% but lower precision of 91.67%. The weighted-average ensemble, on the other hand, performed better than both models. It had an accuracy of 98.18% and an F1-score of 97.78%. It also maintained a recall of 100.00% and increased precision to 95.65%. In this sense, the WA-Ensemble preserves the high ASD sensitivity observed in VGG16 while improving the overall balance of the classification, particularly through increased precision and F1-score. These findings confirm that the ensemble produces more accurate and robust predictions than the individual models.

To provide statistical validation of the reported results, we estimated 95% confidence intervals (CIs) for the evaluation metrics using bootstrap resampling and report them in Table 3. We generated 1000 bootstrap samples from the test set by sampling with replacement. For each resampled test set, we computed accuracy, precision, recall, F1-score,

and ROC-AUC. We obtained the lower and upper bounds of each 95% CI from the 2.5th and 97.5th percentiles of the corresponding bootstrap distribution. These intervals provide an estimate of the uncertainty associated with each performance metric and support a more robust interpretation of the reported improvements.

Table 3. An overview of the performance comparison between the WA-Ensemble and individual models, including Xception and VGG16, based on accuracy, precision, recall, and F1-score (with 95% confidence intervals shown in parentheses). The best results per metric have been underlined and highlighted in bold.

Metric	Xception	VGG16	WA-Ensemble
Accuracy	0.9636 (0.9091–1.0000)	0.9636 (0.9091–1.0000)	<u>0.9818</u> (0.9455–1.0000)
Precision	0.9545 (0.8421–1.0000)	0.9167 (0.7826–1.0000)	<u>0.9565</u> (0.8421–1.0000)
Recall	0.9545 (0.8421–1.0000)	<u>1.0000</u> (1.0000–1.0000)	<u>1.0000</u> (1.0000–1.0000)
F1-score	0.9545 (0.8718–1.0000)	0.9565 (0.8780–1.0000)	<u>0.9778</u> (0.9143–1.0000)
ROC-AUC	0.9862 (0.9560–1.0000)	0.9945 (0.9769–1.0000)	<u>0.9959</u> (0.9800–1.0000)

For clarity, we also show the results of the WA-Ensemble using a confusion matrix as depicted in Figure 5. As we can observe, the WA-Ensemble achieves strong, reliable classification performance. With our approach, all 22 ASD cases are correctly identified, with no false negatives. Regarding the TD cases, 32 are correctly identified, with only 1 misclassified as ASD, yielding a very low false-positive rate. The results show perfect recall for the ASD class, ensuring that all the ASD cases are detected. Overall, the confusion matrix results suggest that the WA-Ensemble achieves high performance while maintaining an excellent balance between sensitivity and specificity, making it well-suited for robust ASD classification.

		WA Ensemble	
		TD	ASD
Actual class	TD	32 58.18%	1 1.81%
	ASD	0 0.00%	22 40.00%
		Predicted class	

Figure 5. Confusion matrix of the WA-Ensemble approach for ASD detection on the test set.

The ROC-AUC curves shown in Figure 6 exhibit the performance of the individual models Xception and VGG16, as well as the WA-Ensemble for ASD classification. The Xception model achieved an ROC-AUC value of 98.62%, while VGG16 further improved to 99.45%. The WA-Ensemble outperformed both models with a higher ROC-AUC of 99.59%, offering the most accurate, reliable classification with improved generalization and robustness compared to individual models.

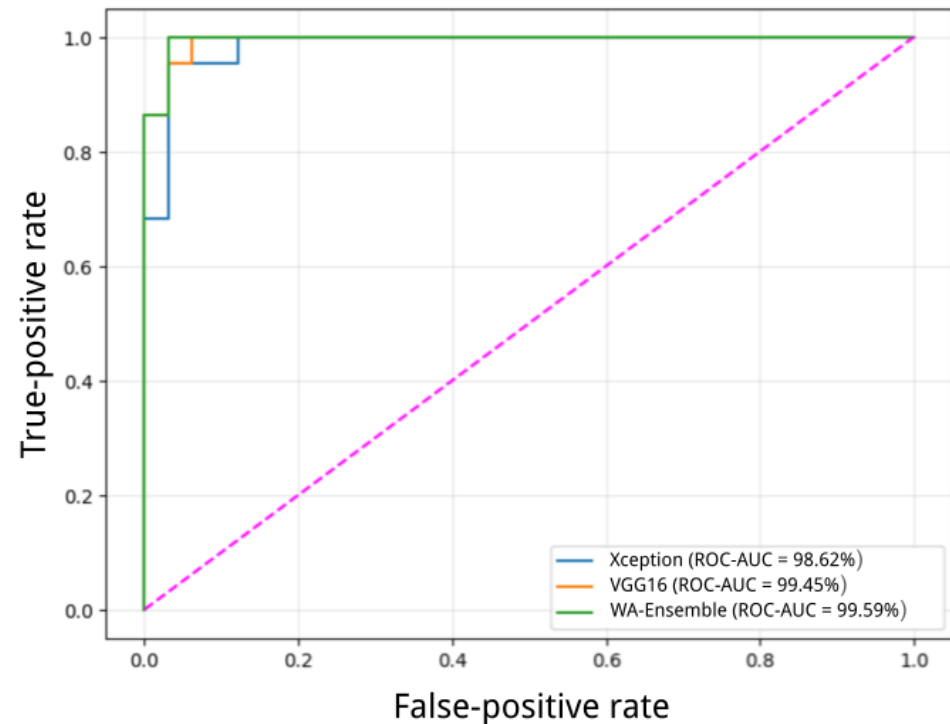


Figure 6. ROC-AUC curves of the baseline models (Xception and VGG16) and the WA-Ensemble for ASD detection.

We show the class-wise performance of the individual DCNN models, Xception and VGG16, and the WA-Ensemble for ASD detection using the VETSP image dataset in Table 4. The Xception model produces identical values for the three metrics analyzed (precision, recall, and F1-score, all with 96.97%) for the TD class. Conversely, for the ASD class, it achieves 95.45% across the three metrics. This consistency shows that the model can reliably distinguish between classes without significant bias toward any class. Xception demonstrated reliable generalization power, with an overall weighted average performance of 96.36% across all parameters. The VGG16 model, on the other hand, shows complementary behavior. VGG16 achieves perfect precision (100.00%) but a relatively low recall (93.93%) for the TD class, suggesting conservative predictions with fewer false positives. The model achieves perfect recall (100.00%) but lower precision (91.66%) for the ASD class, indicating a tendency to prioritize sensitivity to ASD patients at the expense of more false positives. The weighted average F1-score of 96.38% remains competitive despite this imbalance. The combined strengths of the Xception and VGG16 models are effectively harnessed through a weighted-average ensemble. The ensemble model demonstrates superior class-specific performance, achieving 100.00% precision and 96.96% recall for the TD class, and 95.65% precision with perfect recall of 100.00% for the ASD class. This balanced behavior results in the highest F1-scores for both classes: 98.46% for TD and 97.78% for ASD. Overall, the ensemble achieves a weighted average performance of 98.26% precision, 98.18% recall, and 98.18% F1-score, surpassing the performance of both individual models. These findings demonstrate that combining the complementary prediction features of separate DCNN

models via a weighted-average ensemble improves robustness and generalization, resulting in better and more reliable ASD detection performance.

Table 4. Class-wise performance evaluation of individual models, including Xception, VGG16, and the WA-Ensemble, based on evaluation metrics, precision, recall, F1-score, and support (number of samples per class). The best results per model and metric have been underlined and highlighted in bold.

Model	Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Xception	TD	<u>96.97</u>	<u>96.97</u>	<u>96.97</u>	33
	ASD	95.45	95.45	95.45	22
	Weighted average	96.36	96.36	96.36	55
VGG16	TD	<u>100.00</u>	93.93	<u>96.87</u>	33
	ASD	91.67	<u>100.00</u>	95.65	22
	Weighted average	96.66	96.36	96.38	55
WA-Ensemble	TD	<u>100.00</u>	96.96	98.46	33
	ASD	95.65	<u>100.00</u>	97.78	22
	Weighted average	98.26	98.18	98.18	55

4.3. Effect of Data Augmentation on Individual Models and Ensemble Performance

To explicitly assess the effect of data augmentation, Table 5 compares the performance of each model with and without augmentation. As shown, data augmentation improved the performance of both individual models and the WA-Ensemble, suggesting better learning and generalization. The WA-Ensemble integrates complementary model predictions, combining their strengths for improved results. The ensemble with data augmentation achieves the best results, demonstrating superior resilience and classification performance, compared with individual models and non-augmented configurations.

Table 5. A comparative analysis of each model (Xception, VGG16, and WA-Ensemble), with and without data augmentation, on the VETSP image dataset using accuracy, precision, recall, and F1-score. The best results per model and metric have been underlined and highlighted in bold.

Model	Augmentation	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Xception	No	92.73	95.00	86.36	90.48
	Yes	<u>96.36</u>	<u>95.45</u>	<u>95.45</u>	<u>95.45</u>
VGG16	No	90.91	<u>94.74</u>	81.82	87.80
	Yes	<u>96.36</u>	91.67	<u>100.00</u>	<u>95.65</u>
WA-Ensemble	No	94.55	95.24	90.91	93.02
	Yes	<u>98.18</u>	<u>95.65</u>	<u>100.00</u>	<u>97.78</u>

4.4. Comparative Analysis with State-of-the-Art Pretrained Deep Learning Architectures

In this experiment, we conducted a comprehensive comparative analysis to assess the effectiveness of WA-Ensemble against widely used pretrained DCNN architectures, including EfficientNetB1 [55], EfficientNetB2 [55], EfficientNetB3 [55], EfficientNetB4 [55], ResNet101 [56], ResNet101V2 [57], ResNet50 [56], ResNet50V2 [57], MobileNet [58], MobileNetV2 [59], VGG19 [42], DenseNet121 [60], InceptionResNetV2 [61], InceptionV3 [62], and the proposed approach, WA-Ensemble. Table 6 shows the performance outcomes for all models. The WA-Ensemble demonstrated superior results across accuracy, recall, F1-score, and AUC compared to individual pretrained models. This highlights the advantage of combining multiple complementary DCNN models via an ensemble strategy, reducing error rates, mitigating overfitting, and improving generalization and robustness

in ASD classification on ETSP images. The findings suggest that the employed approach effectively captures features that individual models may miss, providing more reliable and discriminative information.

Table 6. Performance comparison of relevant DL Models from the literature and WA-Ensemble for ASD Classification on the test set. The best results per metric have been underlined and highlighted in bold.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
EfficientNetB1	90.91	<u>100.00</u>	77.27	87.18	95.45
EfficientNetB2	92.73	<u>100.00</u>	81.82	90.00	95.45
EfficientNetB3	90.91	94.74	81.82	87.80	94.21
EfficientNetB4	94.55	91.30	95.45	93.33	94.90
ResNet101	92.73	87.50	95.45	91.30	96.69
ResNet101V2	94.55	95.24	90.91	93.02	98.07
ResNet50	94.55	95.24	90.91	93.02	95.87
ResNet50V2	94.55	95.24	90.91	93.02	98.62
MobileNet	90.91	81.84	<u>100.00</u>	89.80	96.69
MobileNetV2	89.09	80.77	95.45	87.50	98.76
VGG19	92.73	87.50	95.45	91.30	98.62
DenseNet121	94.55	91.30	95.45	93.33	98.62
InceptionResNetV2	87.27	89.47	77.27	82.93	97.52
InceptionV3	89.09	86.36	86.36	86.36	96.01
Xception	96.36	95.45	95.45	95.45	98.62
VGG16	96.36	91.67	<u>100.00</u>	95.65	99.45
WA-Ensemble	<u>98.18</u>	95.65	<u>100.00</u>	<u>97.78</u>	<u>99.59</u>

4.5. Benchmarking Against Existing ASD Detection Methods Using the VETSP Image Dataset

The approach proposed in this work, WA-Ensemble, uses a weighted-average ensemble that combines the complementary predictive behaviors of Xception and VGG16. Ensemble weights are adjusted via grid search based on the validation set. This method, as previously mentioned (Table 5), achieved an accuracy of 98.18%, precision of 95.65%, 100.00% recall, and an F1-score of 97.78%.

However, other recent studies have also shown the effectiveness of approaches that analyze VETSP images for ASD detection. For instance, Alsaïdi et al. [14] introduced the T-CNN-ASD model, which achieved 95.59% accuracy, 77.60% recall, and 78.73% F1-score on the VETSP dataset. These results demonstrate satisfactory classification performance. However, the comparatively lower sensitivity and F1-score may restrict its robustness in clinical screening applications. In a different approach, Mumenin et al. [36] presented ASDNet, an involution-based deep architecture, and reported robust performance across various eye-tracking representations, achieving accuracies of up to 98.12%, along with 97.89% precision, 99.03% recall, and 99.72% AUC, demonstrating a promising discriminative capability. Meanwhile, Jaradat et al. [22] presented a hybrid framework that combines MobileNet feature extraction with stacking EL (SVM and KNN), achieving accuracies of 96.10% to 98.18% across various datasets. However, the absence of F1-score and AUC metrics limits a comprehensive evaluation of classification balance.

We adopted the implementations of these three models and tested them on the same VETSP image dataset. From Table 7, we can observe that WA-Ensemble surpasses the results obtained by the approach of [14] across all metrics and provides more balanced performance than the approach of Jaradat et al. [22]. Besides, it remains competitive with the notable results obtained by the model of Mumenin et al. [36]. Collectively, these results demonstrate that appropriately weighted fusion of heterogeneous CNN architectures can improve generalization and resilience for VETSP image-based ASD detection.

Table 7. Performance comparison between the proposed approach and existing methods on the VETSP image dataset using accuracy, precision, recall, F1-score, and AUC. The best results per metric have been underlined and highlighted in bold.

Reference	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Alsaïdi et al. [14]	95.59	NA	77.60	78.73	NA
Mumenin et al. [36]	98.12	97.89	99.03	98.55	99.72
Jaradat et al. [22]	98.18	98.20	98.00	NA	NA
Proposed approach	98.18	95.65	100.00	97.78	99.59

5. Conclusions

In this study, we evaluated the effectiveness of a weighted-average ensemble approach that combines the complementary predictive strengths of Xception and VGG16 for ASD detection using eye-tracking scanpath images from the VETSP image dataset. The individual evaluation of both models showed that Xception and VGG16 achieved the same overall accuracy of 96.36%, while exhibiting different predictive behaviors. Xception provided a more balanced trade-off between precision and recall, whereas VGG16 achieved perfect recall for ASD but showed a higher tendency to over-classify ASD cases. These complementary error patterns support the motivation for integrating both models through a weighted-average ensemble rather than relying on either architecture individually.

The proposed WA-Ensemble effectively leveraged these complementary behaviors by combining the prediction probabilities of both models using validation-optimized weights. As a result, the ensemble achieved 98.18% accuracy, 95.65% precision, 100.00% recall, 97.78% F1-score, and 99.59% ROC-AUC. The model correctly identified all ASD cases in the test set while maintaining a low false-positive rate, which is particularly relevant for screening-oriented applications where sensitivity is a key requirement. In addition, the bootstrap-based confidence intervals provided an estimate of the uncertainty associated with the reported metrics, supporting a more robust interpretation of the results.

The class-wise analysis further confirmed the benefit of the ensemble strategy. The WA-Ensemble achieved strong performance for both ASD and TD classes, suggesting that weighted fusion helped reduce the limitations of the individual architectures and produced a more balanced classification behavior. The use of data augmentation also improved the performance of the individual models and the ensemble, indicating that augmentation and weighted fusion can work complementarily when dealing with small eye-tracking image datasets.

Compared with the pretrained architectures evaluated in this study, the WA-Ensemble achieved the best overall performance across accuracy, recall, F1-score, and ROC-AUC. When compared with existing ASD detection methods evaluated on the VETSP dataset, the proposed approach remained highly competitive with the best reported results, particularly due to its perfect recall and strong overall balance between performance, simplicity, and interpretability. These findings suggest that validation-optimized weighted fusion of complementary CNN architectures can provide a practical and reliable alternative for ASD detection using eye-tracking scanpath images, without requiring more complex ensemble mechanisms such as stacking, boosting, or repeated resampling procedures.

Clinically, the high recall achieved by the WA-Ensemble is relevant for screening-oriented applications, where minimizing false negatives is important to avoid delaying referral and intervention for children with ASD. The model correctly identified all ASD cases in the test set while producing only one false positive among TD cases, suggesting that it may support clinicians in identifying children who require further evaluation. Nevertheless, the proposed approach should be regarded as a decision-support tool rather

than a standalone diagnostic system, and further validation on larger and independent datasets is required before clinical use.

Despite the promising results, some limitations should be acknowledged. First, the VETSP dataset is relatively small. Although data augmentation and ensemble learning improved performance under the adopted experimental protocol, future studies should validate the proposed approach on larger, multicenter datasets. Second, because this study used an image-level split, images from the same participant may appear in different subsets, introducing shared subject-specific characteristics that could lead to data leakage and optimistic performance estimates. For this reason, future work should also evaluate the method under a participant-level split to provide a stricter assessment of generalization to unseen individuals. Finally, incorporating explainability mechanisms would help clarify the visual patterns driving the model's predictions and increase the transparency and clinical trustworthiness of the proposed approach.

Author Contributions: Conceptualization, M.A. and J.C.O.-B.; methodology, M.A., S.H., I.A. and J.C.O.-B.; software, M.A. and J.C.O.-B.; validation, M.A., S.H., I.A. and J.C.O.-B.; formal analysis, M.A., I.A. and J.C.O.-B.; investigation, M.A. and J.C.O.-B.; resources, M.A. and J.C.O.-B.; data curation, M.A., S.H. and J.C.O.-B.; writing—original draft preparation, M.A., S.H., I.A. and J.C.O.-B.; writing—review and editing, M.A., S.H., I.A. and J.C.O.-B.; visualization, M.A., S.H. and J.C.O.-B.; supervision, M.A., I.A. and J.C.O.-B.; project administration, M.A. and J.C.O.-B.; funding acquisition, M.A. and J.C.O.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available in FigShare at <https://doi.org/10.6084/m9.figshare.7073087>.

Acknowledgments: During the preparation of this document, the authors used ChatGPT (GPT-5, OpenAI) to develop and edit the manuscript, improving its consistency, clarity, and linguistic structure. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

ADI	Autism Diagnostic Interview
ADOS	Autism Diagnostic Observation Schedule
AI	Artificial Intelligence
ANN	Artificial Neural Network
ASD	Autism Spectrum Disorder
AUC	Area Under the Curve
BDT	Boosted Decision Tree
CDC	Centers for Disease Control and Prevention
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DJ	Decision Jungle
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
EEG	Electroencephalogram

EL	Ensemble Learning
ETSP	Eye Tracking Scanpath
GB	Gradient Boosting
KNN	K-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NB	Naive Bayes
PCA	Principal Component Analysis
RF	Random Forest
SVM	Support Vector Machine
WHO	World Health Organization
XGB	Extreme Gradient Boosting

References

1. Olusanya, B.O.; Storbeck, C.; Cheung, V.G.; Hadders-Algra, M. Disabilities in early childhood: A global health perspective. *Children* **2023**, *10*, 155. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Zeidan, J.; Fombonne, E.; Scorah, J.; Ibrahim, A.; Durkin, M.S.; Saxena, S.; Yusuf, A.; Shih, A.; Elsabbagh, M. Global prevalence of autism: A systematic review update. *Autism Res.* **2022**, *15*, 778–790. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Kanner, L. Autistic disturbances of affective contact. *Nerv. Child* **1943**, *2*, 217–250.
4. Barahona-Corrêa, J.; Filipe, C.N. A concise history of Asperger syndrome: The short reign of a troublesome diagnosis. *Front. Psychol.* **2016**, *6*, 2024. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Hodges, H.; Fealko, C.; Soares, N. Autism spectrum disorder: Definition, epidemiology, causes, and clinical evaluation. *Transl. Pediatr.* **2020**, *9*, S55. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Baj, J.; Flieger, W.; Flieger, M.; Forma, A.; Sitarz, E.; Skórzyńska-Dzidusko, K.; Grochowski, C.; Maciejewski, R.; Karakuła-Juchnowicz, H. Autism spectrum disorder: Trace elements imbalances and the pathogenesis and severity of autistic symptoms. *Neurosci. Biobehav. Rev.* **2021**, *129*, 117–132. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Volkmar, F.R.; Reichow, B.; McPartland, J. Classification of autism and related conditions: Progress, challenges, and opportunities. *Dialogues Clin. Neurosci.* **2012**, *14*, 229–237. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Wankhede, N.L.; Kale, M.B.; Shukla, M.; Nathiya, D.; Roopashree, R.; Kaur, P.; Goyanka, B.; Rahangdale, S.R.; Taksande, B.G.; Upaganlawar, A.B.; et al. Leveraging AI for the diagnosis and treatment of autism spectrum disorder: Current trends and future prospects. *Asian J. Psychiatry* **2024**, *101*, 104241. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Vakadkar, K.; Purkayastha, D.; Krishnan, D. Detection of autism spectrum disorder in children using machine learning techniques. *SN Comput. Sci.* **2021**, *2*, 1–9. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Zwaigenbaum, L.; Bauman, M.L.; Choueiri, R.; Kasari, C.; Carter, A.; Granpeesheh, D.; Mailloux, Z.; Smith Roley, S.; Wagner, S.; Fein, D.; et al. Early intervention for children with autism spectrum disorder under 3 years of age: Recommendations for practice and research. *Pediatrics* **2015**, *136*, S60–S81. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Uddin, M.Z.; Shahriar, M.A.; Mahamood, M.N.; Alnajjar, F.; Pramanik, M.I.; Ahad, M.A.R. Deep learning with image-based autism spectrum disorder analysis: A systematic review. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107185. [\[CrossRef\]](#)
12. Setu, D.M. A Systematic Review of Developments in Eye Tracking and Machine Learning for the Early Detection of Autism Spectrum Disorder. *J. Technol. Behav. Sci.* **2025**, *10*, 793–810. [\[CrossRef\]](#)
13. Kanhirakadavath, M.R.; Chandran, M.S.M. Investigation of eye-tracking scan path as a biomarker for autism screening using machine learning algorithms. *Diagnostics* **2022**, *12*, 518. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Alsaidi, M.; Obeid, N.; Al-Madi, N.; Hiary, H.; Aljarah, I. A convolutional deep neural network approach to predict autism spectrum disorder based on eye-tracking scan paths. *Information* **2024**, *15*, 133. [\[CrossRef\]](#)
15. Al Kuwaiti, A.; Nazer, K.; Al-Reedy, A.; Al-Shehri, S.; Al-Muhanna, A.; Subbarayalu, A.V.; Al Muhanna, D.; Al-Muhanna, F.A. A review of the role of artificial intelligence in healthcare. *J. Pers. Med.* **2023**, *13*, 951. [\[CrossRef\]](#)
16. Kollias, K.F.; Syriopoulou-Delli, C.K.; Sarigiannidis, P.; Fragulis, G.F. The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A systematic review. *Electronics* **2021**, *10*, 2982. [\[CrossRef\]](#)
17. Kalota, F. A primer on generative artificial intelligence. *Educ. Sci.* **2024**, *14*, 172. [\[CrossRef\]](#)
18. Fahim, Y.A.; Hasani, I.W.; Kabba, S.; Ragab, W.M. Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives. *Eur. J. Med. Res.* **2025**, *30*, 848. [\[CrossRef\]](#) [\[PubMed\]](#)

19. Huda, S.; Khan, D.M.; Masroor, K.; Rashid, A.; Shabbir, M. Advancements in automated diagnosis of autism spectrum disorder through deep learning and resting-state functional mri biomarkers: A systematic review. *Cogn. Neurodyn.* **2024**, *18*, 3585–3601. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Syriopoulou-Delli, C.K. Advances in Autism Spectrum Disorder (ASD) diagnostics: From theoretical frameworks to ai-driven innovations. *Electronics* **2025**, *14*, 951. [\[CrossRef\]](#)
21. Shaikh, T.A.; Rasool, T.; Verma, P.; Mir, W.A. A fundamental overview of ensemble deep learning models and applications: Systematic literature and state of the art. *Ann. Oper. Res.* **2024**. [\[CrossRef\]](#)
22. Jaradat, A.S.; Wedyan, M.; Alomari, S.; Barhoush, M.M. Using Machine Learning to Diagnose Autism Based on Eye Tracking Technology. *Diagnostics* **2024**, *15*, 66. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Rawat, W.; Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput.* **2017**, *29*, 2352–2449. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Dhillon, A.; Verma, G.K. Convolutional neural network: A review of models, methodologies and applications to object detection. *Prog. Artif. Intell.* **2020**, *9*, 85–112. [\[CrossRef\]](#)
25. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A. Ensemble learning for disease prediction: A review. *Healthcare* **2023**, *11*, 1808. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Yang, Y.; Lv, H.; Chen, N. A survey on ensemble learning under the era of deep learning. *Artif. Intell. Rev.* **2023**, *56*, 5545–5589. [\[CrossRef\]](#)
27. Khan, A.A.; Chaudhari, O.; Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* **2024**, *244*, 122778. [\[CrossRef\]](#)
28. Carette, R.; Elbattah, M.; Dequen, G.; Guérin, J.L.; Cilia, F. Visualization of eye-tracking patterns in autism spectrum disorder: method and dataset. In *Proceedings of the 2018 Thirteenth International Conference on Digital Information Management (ICDIM)*; IEEE: New York, NY, USA, 2018; pp. 248–253. [\[CrossRef\]](#)
29. Carette, R.; Elbattah, M.; Cilia, F.; Dequen, G.; Guerin, J.L.; Bosche, J. Learning to Predict Autism Spectrum Disorder based on the Visual Patterns of Eye-tracking Scanpaths. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)—HEALTHINF*; SciTePress: Setúbal, Portugal, 2019; pp. 103–112. [\[CrossRef\]](#)
30. Akter, T.; Ali, M.H.; Khan, M.I.; Satu, M.S.; Moni, M.A. Machine learning model to predict autism investigating eye-tracking dataset. In *Proceedings of the 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*; IEEE: New York, NY, USA, 2021; pp. 383–387. [\[CrossRef\]](#)
31. Li, M.; Tang, D.; Zeng, J.; Zhou, T.; Zhu, H.; Chen, B.; Zou, X. An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. *Comput. Speech Lang.* **2019**, *56*, 80–94. [\[CrossRef\]](#)
32. Elbattah, M.; Carette, R.; Dequen, G.; Guérin, J.L.; Cilia, F. Learning clusters in autism spectrum disorder: Image-based clustering of eye-tracking scanpaths with deep autoencoder. In *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; IEEE: New York, NY, USA, 2019; pp. 1417–1420. [\[CrossRef\]](#)
33. Cilia, F.; Carette, R.; Elbattah, M.; Dequen, G.; Guérin, J.L.; Bosche, J.; Vandromme, L.; Le Driant, B. Computer-aided screening of autism spectrum disorder: Eye-tracking study using data visualization and deep learning. *JMIR Hum. Factors* **2021**, *8*, e27706. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Kabir Mehedi, M.H.; Arafin, I.; Md Hasib, K.; Rahman, F.; Alam, M.M.; Tasin, R.; Alim Rasel, A. Early autism disorder detection through visualizing eye-tracking patterns using compact convolutional transformers. In *Proceedings of the 9th International Conference on Computer Technology Applications*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 109–114. [\[CrossRef\]](#)
35. Duan, H.; Min, X.; Fang, Y.; Fan, L.; Yang, X.; Zhai, G. Visual attention analysis and prediction on human faces for children with autism spectrum disorder. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2019**, *15*, 1–23. [\[CrossRef\]](#)
36. Mumenin, N.; Yousuf, M.A.; Nashiry, M.A.; Azad, A.; Alyami, S.A.; Lio', P.; Moni, M.A. ASDNet: A robust involution-based architecture for diagnosis of autism spectrum disorder utilising eye-tracking technology. *IET Comput. Vis.* **2024**, *18*, 666–681. [\[CrossRef\]](#)
37. Soloh, R.; Abou Orm, L.; Dabdoub, D. An image-based analytics framework for early autism detection using eye movements. *Healthc. Anal.* **2025**, *8*, 100439. [\[CrossRef\]](#)
38. Ahmed, Z.A.; Aldhyani, T.H.; Alzahrani, E.M.; Albalawi, E.; Algarni, M.H.; Jadhav, M.E.; Ahmad, S.; Al-Adhaileh, M.H.; Alsubari, S.N.; Mehdi, A. Effectiveness of Histogram Equalization and Ensemble Deep Learning Techniques for Detecting Autism Using Eye-Tracking. *SN Comput. Sci.* **2026**, *7*, 245. [\[CrossRef\]](#)
39. Mousli, S.; Taheri, S.; He, E. ConASD: Contrastive Few Shot Learning for Detecting Autism Spectrum Disorder via Eye Tracking Scanpath: S. Mousli et al. *Multimed. Syst.* **2025**, *31*, 312. [\[CrossRef\]](#)
40. Duan, H.; Zhai, G.; Min, X.; Che, Z.; Fang, Y.; Yang, X.; Gutiérrez, J.; Callet, P.L. A dataset of eye movements for the children with autism spectrum disorder. In *Proceedings of the 10th ACM Multimedia Systems Conference*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 255–260. [\[CrossRef\]](#)

41. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 1251–1258.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [[CrossRef](#)]
43. Sakib, M.; Mustajab, S.; Alam, M. Ensemble deep learning techniques for time series analysis: A comprehensive review, applications, open issues, challenges, and future directions. *Clust. Comput.* **2025**, *28*, 73. [[CrossRef](#)]
44. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [[CrossRef](#)]
45. Ganaie, M.A.; Hu, M.; Malik, A.K.; Tanveer, M.; Suganthan, P.N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [[CrossRef](#)]
46. Khwaja, A.; Naeem, M.; Anpalagan, A.; Venetsanopoulos, A.; Venkatesh, B. Improved short-term load forecasting using bagged neural networks. *Electr. Power Syst. Res.* **2015**, *125*, 109–115. [[CrossRef](#)]
47. Zhang, W.; Jiang, J.; Shao, Y.; Cui, B. Snapshot boosting: A fast ensemble framework for deep neural networks. *Sci. China Inf. Sci.* **2020**, *63*, 112102. [[CrossRef](#)]
48. Bihri, H.; Charaf, L.A.; Azzouzi, S.; Charaf, M.E.H. A Robust Stacking-Based Ensemble Model for Predicting Cardiovascular Diseases. *AI* **2025**, *6*, 160. [[CrossRef](#)]
49. Li, J.; Chang, H.; Yang, J.; Luo, W.; Fu, Y. Visual representation and classification by learning group sparse deep stacking network. *IEEE Trans. Image Process.* **2017**, *27*, 464–476. [[CrossRef](#)] [[PubMed](#)]
50. Chakraborty, G.S.; Batra, S.; Singh, A.; Muhammad, G.; Torres, V.Y.; Mahajan, M. A novel deep learning-based classification framework for COVID-19 assisted with weighted average ensemble modeling. *Diagnostics* **2023**, *13*, 1806. [[CrossRef](#)] [[PubMed](#)]
51. Anand, V.; Gupta, S.; Gupta, D.; Gulzar, Y.; Xin, Q.; Juneja, S.; Shaikh, A. Weighted average ensemble deep learning model for stratification of brain tumor in MRI images. *Diagnostics* **2023**, *13*, 1320. [[CrossRef](#)] [[PubMed](#)]
52. Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 757–774. [[CrossRef](#)]
53. Shahhosseini, M.; Hu, G.; Pham, H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Mach. Learn. Appl.* **2022**, *7*, 100251. [[CrossRef](#)]
54. Fan, Z.; Yu, Z.; Yang, K.; Chen, W.; Liu, X.; Li, G.; Yang, X.; Chen, C.P. Diverse models, united goal: A comprehensive survey of ensemble learning. *CAAI Trans. Intell. Technol.* **2025**, *10*, 959–982. [[CrossRef](#)]
55. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*; PMLR: Long Beach, CA, USA, 2019; pp. 6105–6114.
56. 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*; IEEE: New York, NY, USA, 2016; pp. 770–778.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
58. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
59. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 4510–4520.
60. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 4700–4708.
61. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2017; Volume 31.
62. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 2818–2826.

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