Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Comparative Study of Lightweight Uncertainty-Aware Bayesian CNNs and Ensemble Models for MRI-Based ASD Diagnosis

Soleen Esmael

soleen.ismael@staff.uoz.edu.krd

University of Zakho

Masoud Hassan

University of Zakho

Method Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9396884/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title: A Comparative Study of Lightweight Uncertainty-Aware Bayesian CNNs and Ensemble Models for MRI-Based ASD Diagnosis

Soleen Hishyer Esmael[1*], Masoud Muhammed Hassan[1, 2, 3]
[1] Department of Computer Science, College of Science, University of Zakho, Duhok, Kurdistan Regain, Iraq
[2] Department of Computer Science, College of Science, Knowledge University, Erbil, 44001, Iraq
[3] PRIME Lab, Scientific Research Center, University of Zakho, Kurdistan Region, Iraq
* Corresponding author: soleen.ismael@staff.uoz.edu.krd

Abstract

Background: Early and accurate diagnosis of Autism Spectrum Disorder (ASD) is essential for improving clinical outcomes. Although deep learning (DL) has demonstrated strong potential in analyzing structural Magnetic Resonance Imaging (sMRI), most existing frameworks rely on deterministic models that lack uncertainty quantification—an essential component for clinical decision-making—and often involve high computational complexity, limiting real-world applicability.

Objective: This study aims to conduct a comprehensive comparative evaluation of three deep learning paradigms to develop a robust and uncertainty-aware framework for ASD diagnosis using multi-orientational MRI data (axial, coronal, and sagittal views).

Methods: We systematically evaluated three categories of models: **(1)** fine-tuned deterministic architectures, including EfficientNetV2B0, EfficientNetB3, and MobileNetV2; **(2)** ensemble learning strategies, including averaging, majority voting, weighted averaging, and stacked generalization; and **(3)** lightweight Bayesian convolutional neural networks (CNNs) incorporating Monte Carlo (MC) Dropout and Variational Inference (VI) to estimate predictive uncertainty. This hierarchical framework enables a balanced assessment of diagnostic accuracy and model reliability.

Results: Experimental results demonstrate that ensemble models achieved the highest classification performance, with a peak accuracy of 99.51% in the coronal view. Notably, the proposed lightweight Bayesian CNNs delivered competitive performance while providing well-calibrated uncertainty estimates, enabling the identification of ambiguous and out-of-distribution cases. Additionally, Grad-CAM visualizations revealed that model attention is concentrated in clinically relevant regions, particularly the frontal and temporal lobes, supporting the neurobiological plausibility of the predictions.

Conclusion: This study highlights that integrating uncertainty quantification into lightweight deep learning architectures enhances both transparency and reliability compared to conventional deterministic approaches.

Keywords: Uncertainty Quantification, Ensemble approaches, Explainability Artificial Intelligence, Pre-trained model.

1 Introductions

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental condition characterized by persistent deficits in social communication, behavioral flexibility, and adaptive functioning (Lin et al. 2025). Given the significant neuroplasticity of the developing brain, early and precise diagnosis is paramount for initiating targeted interventions that optimize long-term developmental outcomes. While clinical diagnosis traditionally relies on subjective behavioral observations, structural Magnetic Resonance Imaging (sMRI) has emerged as a critical non-invasive modality for identifying objective neuroanatomical biomarkers. The high-dimensional spatial information captured by MRI enables the detection of subtle cortical variations and volumetric anomalies, making it a premier candidate for sophisticated computational analysis.

In recent years, Deep Learning (DL) has revolutionized medical image analysis by automating the extraction of complex hierarchical features, bypassing the limitations of manual feature engineering. Specifically, Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performance in classifying various neurological disorders (LeCun, Bengio, and Hinton 2015). To mitigate the challenges of limited clinical datasets, transfer learning—utilizing pre-trained architectures fine-tuned on neuroimaging data—has become a standard paradigm to enhance generalizability and convergence stability (Salehi et al. 2023).

Despite these advancements, the clinical adoption of DL models is often hindered by their deterministic nature. Conventional CNNs provide point-estimate predictions that lack a measure of (confidence) or uncertainty. In high-stakes clinical environments, a (black-box) prediction without quantified reliability can lead to misplaced trust or catastrophic misdiagnosis. Furthermore, many existing models are (heavyweight), demanding prohibitive computational resources that are often unavailable in real-world clinical settings. Additionally, the extreme heterogeneity of ASD brain morphology often results in inconsistent performance across different MRI orientations (Axial, Coronal, and Sagittal), a challenge that monolithic, deterministic models struggle to address.

To surmount these barriers, this study proposes a comprehensive comparative framework that evaluates three synergistic computational pillars. **First**, Ensemble Learning strategies (Ganaie et al. 2022)— including average voting, weighted averaging, stacking, and max voting—combine predictions from multiple models to reduce individual biases, enhance robustness, and improve MRI-based ASD classification accuracy. **Second**, Bayesian Deep Learning (BDL), employing Monte Carlo (MC) Dropout or Variational implements Monte Carlo (MC) Dropout and Variational Inference (VI) to transform deterministic outputs into probabilistic distributions; this enables Uncertainty Quantification (UQ), a vital requirement for establishing clinical trust. (Goan and Fookes 2020). Complementing these approaches, explainable AI (XAI) techniques, particularly Gradient-weighted Class Activation Mapping (Grad-CAM) (Panwar et al. 2020), leverages Gradient-weighted Class Activation Mapping (Grad-CAM) to provide interpretable visualizations of the neuroanatomical regions—specifically the frontal and temporal lobes—driving the model's decision, thereby fostering transparency and evidence-based clinical adoption.

1.1 Motivation and Problem Statement

Although existing studies have reported high classification accuracy in ASD diagnosis using DL models, many of these approaches prioritize predictive performance at the expense of clinical interpretability, reliability, and computational efficiency. In particular, most current frameworks lack mechanisms to quantify uncertainty, limiting their ability to identify ambiguous or out-of-distribution cases—an essential requirement for safe clinical deployment. Moreover, the absence of systematic evaluations across

multiple MRI orientations restricts the understanding of view-specific diagnostic relevance and robustness.

Motivated by these challenges, this study aims to bridge the gap between high-performance predictive modeling and clinically reliable, interpretable AI systems. Specifically, a comprehensive comparative analysis is conducted to investigate whether ensemble-based approaches, which enhance predictive accuracy, or Bayesian methods, which provide uncertainty-aware predictions, offer a more effective and practical solution for ASD diagnosis in real-world clinical settings, particularly under resource constraints.

1.2 Contributions and Objectives

The primary objective of this research is to architect and systematically compare a suite of lightweight, uncertainty-aware diagnostic frameworks for ASD. The specific contributions of this work are as follows

- **Multi-View Benchmarking:** Implementation of an optimized transfer learning pipeline using EfficientNetV2B0, EfficientNetB3, and MobileNet to establish a performance baseline across all three MRI orientations.
- **Robust Ensemble Integration:** Development of high-order ensemble strategies to maximize diagnostic sensitivity and specificity while reducing variance.
- **Probabilistic Modeling:** Design of Lightweight Bayesian CNNs and the adaptation of pre-trained models via Bayesian Inference to integrate predictive uncertainty into the diagnostic workflow.
- **Neuroanatomical Validation:** Application of Grad-CAM to ensure the model's (attention) aligns with established neurological ASD indicators, fostering clinical trust and biological plausibility.
- **Comparative SOTA Analysis:** A rigorous evaluation demonstrating that our proposed lightweight, uncertainty-aware frameworks achieve superior or competitive performance relative to significantly more complex, resource-intensive state-of-the-art models.

The remainder of this article is organized as follows: Section 2 reviews relevant literature in DL-based ASD detection. Section 3 details the proposed methodology, including dataset stratification, Bayesian implementation, and ensemble configurations. Section 4 presents a comparative analysis of experimental results and clinical interpretations. Finally, Section 5 concludes the study and outlines future research trajectories.

1.3 Related Works

Magnetic Resonance Imaging (MRI) has emerged as a powerful modality for investigating neurobiological abnormalities associated with ASD. Numerous deep learning-based studies have utilized structural sMRI, functional fMRI, or combined modalities to capture discriminative brain patterns. These approaches commonly rely on data from the ABIDE I and II datasets and employ advanced techniques such as 2D slice extraction, transfer learning, multimodal fusion, and feature interpretability.

For instance, (Darweesh, Salem, and Al-Atabany 2022) proposed a CNN-based framework for ASD classification using a three-stage pipeline: preprocessing 4D rs-fMRI volumes, converting them into 2D slices, and training a CNN to distinguish ASD from typically developing individuals. Utilizing rs-fMRI data from 141 subjects at the University of California Los Angeles (UCLA) site in the ABIDE dataset, the model achieved exceptional accuracy of 99.8% and 98.3%.

Correspondingly, (Feng and Xu 2023) introduced a custom CNN for early ASD diagnosis using rs-fMRI data from the ABIDE I dataset, specifically the New York University (NYU) site. They extracted 22,176 2D samples from preprocessed 4D scans of 126 children (56 ASD, 70 TD) by decomposing each scan into 3D volumes over time and slicing along axial, sagittal, and coronal planes, combined into a three-channel 2D representation. The model achieved 99.39% accuracy using five-fold cross-validation.

In a similar manner, (Herath et al. 2021)introduced a Inception V3 transfer learning-based framework for ASD classification using fMRI data from 138 subjects (69 ASD, 69 controls) of the Kennedy Krieger Institute (KKI) in the ABIDE dataset. They transformed 4D fMRI volumes into 2D epi, stat map, and glass brain images for feature extraction. Two configurations were evaluated: MED1 (ImageNet pre-trained weights) and MED2 (random initialization). MED1 achieved superior accuracies of 98.35% echo planar imaging (epi), 96.76% (stat maps), and 98.24% (glass brain), highlighting the effectiveness of domain-specific 2D projections with transfer learning.

In another study, (Mostafa, Karim, and Hossain 2023) developed a deep learning framework for ASD detection using structural and functional MRI (sMRI and fMRI) images from the ABIDE I and II datasets. They proposed a trial-and-error method to extract informative 2D slices from 4D fMRI data by selecting axial views based on percentage slice positions to ensure consistency across subjects. Transfer learning was applied using ResNet50, VGG-16, and Inception-V4 architectures, further enhanced with a custom "Healthy Block" to reduce overfitting and improve generalization. Their fusion approach combining sMRI and fMRI predictions achieved notable classification accuracies, with a modified ResNet-50 model attaining 84.04% accuracy on ABIDE II and a ResNet-50 with Inception-V4 ensemble reaching 83.22% on ABIDE I.

Furthermore, (Jain et al. 2023) introduced a hybrid deep learning model for ASD detection from sMRI. The approach involved advanced preprocessing to remove non-brain tissues, followed by hybrid segmentation using Fuzzy C-Means and Gaussian Mixture Models, and feature extraction with VGG-16. Classification was performed using a ResNet architecture optimized by the Dwarf Mongoose algorithm, achieving 99.83% accuracy on the ABIDE dataset. However, the study lacked detailed descriptions of data acquisition across imaging sites.

(Saponaro et al. 2024) proposed a joint fusion deep learning model integrating harmonized structural and functional MRI data from the ABIDE I and II datasets for ASD classification. Morphometric and functional connectivity features were extracted using Freesurfer and the Configurable Pipeline for the Analysis of Connectomes (CPAC) pipeline, respectively. The model outperformed single-modality baselines (AUC = 0.78; accuracy = 0.85) by jointly optimizing modality-specific representations. SHAP analysis revealed key discriminative patterns within the Default Mode Network and the Social Brain.

Moreover, (Prasad, G. V Sriramakrishnan, and Diana Jeba Jingle 2023)conducted an advanced deep learning framework for ASD detection using brain MRI data, incorporating a Hybrid Sewing Training Optimization (HSTO) strategy to train the ZFNet architecture. The methodology integrates Sewing Training-Based Optimization (STBO) with Hybrid Leader-Based Optimization (HLBO) to enhance the convergence speed and classification performance. Pre-processing is carried out using a Wiener filter to reduce noise, followed by Region of Interest (ROI) extraction and

pivotal region extraction using the proposed HSTO approach. The ZFNet classifier, optimized with HSTO, demonstrated superior accuracy 95.7% on the ABIDE dataset.

(Parui et al. 2023) Developed a deep learning–based framework for ASD detection from rs-fMRI data. The study used preprocessed ABIDE-I data generated via the CPAC. Functional connectivity networks were constructed using Pearson correlation across multiple brain atlases. The Low Estimated Rank Tensor (LERT) method was applied to model comparable but non-identical connectivity topologies. These refined networks were classified using 2D-CNN with results from different atlases combined through majority voting. The approach achieved an accuracy of 84.79%.

Although DL has made significant progress in detecting ASD, several important limitations and gaps remain in current research. Many studies have achieved high accuracy using MRI; however, their models are often overly complex, which limits clinical applicability and increases computational demands. Moreover, critical aspects such as quantifying predictive uncertainty—essential for evaluating model confidence and supporting clinical decisions—have been largely neglected. Addressing uncertainty is therefore vital for improving the reliability and interpretability of ASD detection models, especially when paired with lightweight architectures that are more practical for clinical use. To overcome these challenges, our study proposes a Bayesian lightweight model that delivers high performance while maintaining computational efficiency. Additionally, we utilize Grad-CAM to generate visual explanations, enhancing the transparency and interpretability of the model's decision-making process in clinical settings.

2 Methodology

In this study, we propose a robust, multi-stage methodology for ASD classification using structural MRI data, as illustrated in Figure 1. The framework is designed to transition from deterministic deep learning to uncertainty-aware Bayesian inference, ensuring both high diagnostic accuracy and clinical reliability. The workflow is partitioned into three distinct experimental phases:

- **Phase I**: Deterministic Feature Learning: Fine-tuning of state-of-the-art pretrained CNNs, including EfficientNetV2B0, EfficientNetB3, and MobileNet.
- **Phase II**: Ensemble Generalization: Integration of multiple models via average, majority, weighted voting, and stacked generalization to enhance predictive stability.
- **Phase III:** Bayesian Uncertainty Quantification: Implementation of lightweight Bayesian Convolutional Neural Networks (BCNNs) utilizing Monte Carlo (MC) Dropout and Variational Inference (VI) to quantify epistemic and aleatoric uncertainties.

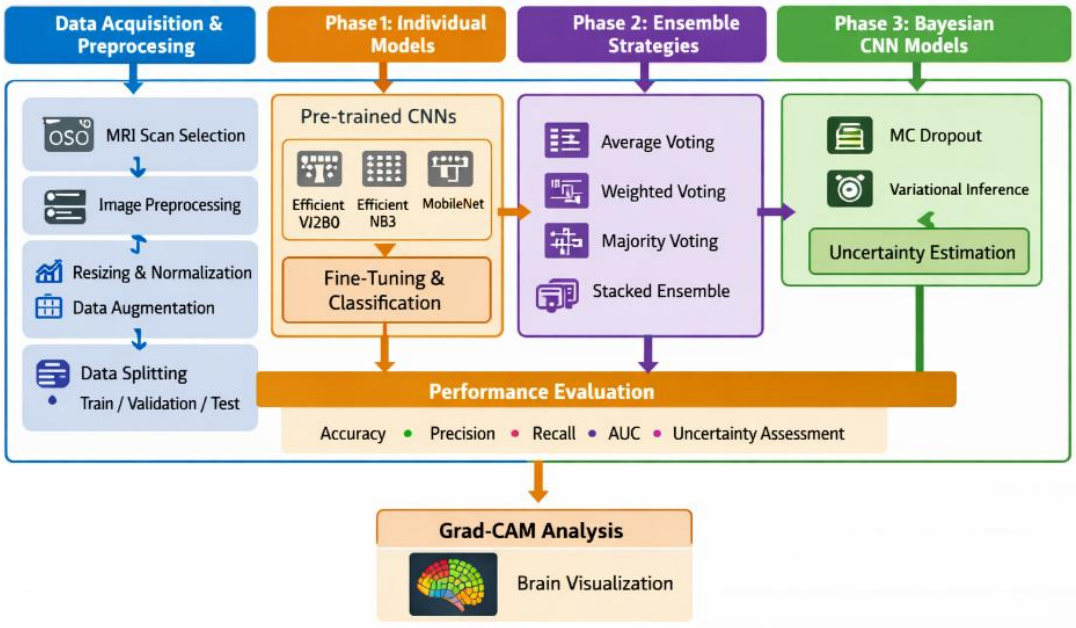


Figure 1: Flow Diagram of the proposed methodology.

2.1 Data Acquisition and Preprocessing

The proposed framework was evaluated using the Autism Brain Imaging Data Exchange II (ABIDE II) dataset, a large-scale multisite repository of NIfTI-format neuroimaging data. To ensure data diversity and model generalizability, we curated a subset of 1,112 participants (539 ASD and 573 typically developing) across six distinct international sites: Kennedy Krieger Institute, NYU Langone Medical Center, Stanford University, Erasmus University Medical Center, University of Miami, and Georgetown University are summarized in Table 1. To capture the full spatial complexity of neuroanatomical variations, 2D slices were extracted across three orthogonal planes—Axial, Sagittal, and Coronal. From the original 3D MRI volumes (182 × 216 × 182 voxels), 100 high-information slices were selected per subject (indices 50–150 for axial/coronal; 50–170 for sagittal). This process yielded a comprehensive dataset of 40,800 axial, 48,960 sagittal, and 40,800 coronal images as illustrate in Table 1.

Prior to model training, raw MRI data were uniformly resized to 224×224 pixels to standardize input dimensions across axial, sagittal, and coronal views, thereby ensuring consistency and enhancing model efficiency. Subsequently, pixel normalization was applied following model-specific preprocessing protocols implemented via

Keras (Anon 2023): EfficientNetB3 inputs were scaled to the [0, 255] range, EfficientNetV2B0 to [0, 1], MobileNet to [-1, 1], while the lightweight model and EfficientNetB0 adhered to their respective normalization schemes, with the lightweight model normalizing pixel intensities to [0, 1] by dividing by 255 (Eq. 1). To mitigate inherent class imbalances and bolster the framework's invariance to spatial variations, data augmentation via horizontal flipping was employed, effectively diversifying the training distribution.

$$x_{Normalized} = \frac{x_{original}}{255} \quad (1)$$

Following preprocessing, MRI slices were extracted separately from the axial, sagittal, and coronal planes, resulting in 40,800 slices each for axial and coronal views, and 48,690 for the sagittal view. Depending on the modeling approach, two distinct data-splitting schemes were employed: ensemble learning utilized a 70% training, 10% validation, and 20% testing split, whereas individual models—MobileNet, EfficientNetB0, and the lightweight BCNNs architecture—employed an 80% training, 10% validation, and 10% testing split to maximize training data availability. Detailed slice distributions per subset and plane are provided in Table 2 This plane-specific partitioning ensured balanced representation and facilitated rigorous model training and evaluation.

Table 1*:* Participant demographics by site that utilized in this study.

Site	ASD			Normal			All
	#Sub	Sex (M/F)	Age (year)	# Sub	Sex (M/F)	Age (year)	
KKI	56	41/15	8-13	155	99/56	8-13	211
NYU	27	24/3	5.1-8.8	0	0	0	27
SU	21	19/2	8-13	21	19/2	8-13	42
EMC	25	20/5	6-11	27	22/5	6-10	52
UM	13	11/2	7-13	15	11/4	7-13	28
GU	24	13/11	8.1-13.9	24	9/15	8.1-13.9	48

Table 2: Data splits for MRI slices across axial, sagittal, and coronal planes.

MRI data splitting with Ensembling approach				
MRI planes	Training 70%	Validation 10%	Testing 20%	Total
Axial	28560	4080	8160	40800
Sagittal	34272	4869	9792	48960
Coronal	28560	4080	8160	40800
MRI data splitting with BCNN model				
MRI planes	Training 80%	Validation 10%	Testing 10%	Total
Axial	32640	4080	4080	40800
Sagittal	39168	4869	4896	48960
Coronal	32640	4080	4080	40800

2.2 Experiment 1: Architectural Adaptation and Fine-Tuning Strategy

To establish a high-fidelity diagnostic baseline, four Pre-trained (variants EfficientNetV2-B0, EfficientNetB3_1, EfficientNetB3_2, and MobileNet—were systematically reconfigured and fine-tuned to extract discriminative neuroanatomical features from structural MRI data. Leveraging a transfer learning paradigm, all backbones were initialized with ImageNet-pretrained weights and subsequently adapted for binary ASD classification through specific architectural modifications.

For the primary configurations (EfficientNetV2-B0 and EfficientNetB3_1), the original classification heads were replaced with a Global Average Pooling (GAP) layer to generate a condensed representation of high-level features, which were then mapped to the output space via a two-unit dense layer activated by a Softmax function:

$$P(y = i|z) = \left(\frac{e^{z_i}}{\sum e^{z_j}}\right) \quad (2)$$

In contrast, to enhance model regularization and mitigate the risk of overfitting in complex neuroimaging domains, the EfficientNetB3_2 and MobileNet variants incorporated an auxiliary latent dense layer with 128 units and a LeakyReLU activation function ($\alpha = 0.1\ and\ 0.2$, respectively). A critical distinction was maintained between these regularized models: EfficientNetB3_2 integrated a dropout layer (rate = 0.2) to enforce stochastic feature independence, whereas MobileNet served as an ablative control by excluding dropout, thereby allowing for a comparative analysis of explicit regularization on ASD biomarker detection. This unified architectural framework, ensures a balance between computational efficiency and the preservation of spatial features across heterogeneous MRI orientations.

2.3 Experiment 2: Ensemble Learning Methods

To optimize diagnostic reliability and mitigate the inherent predictive variance of individual deterministic backbones, a comprehensive ensemble framework was developed based on the integration of model-specific class probabilities. This phase utilized the class-wise confidence scores from four fine-tuned CNNs: EfficientNetV2-B0 (M_1), EfficientNetB3_1 (M_2), EfficientNetB3_2 (M_3), and MobileNet (M_4). To identify the most synergistic combination of base learners and evaluate the contribution of each architecture, we systematically evaluated 15 distinct configurations categorized into four primary subsets As illustrated in Figure 3:

- Configuration A: {M_1, M_2, M_3}
- Configuration B: {M_1, M_2, M_4}
- Configuration C: {M_2, M_3, M_4}
- Configuration D: (Full Integration): {M_1, M_2, M_3, M_4}

Four fusion paradigms were mathematically formulated to synthesize these outputs into a unified diagnostic decision.

First, the **average voting strategy** (implemented in four configurations) computes the arithmetic mean of the predicted probabilities generated by the selected models. This approach smooths the decision boundary by mitigating model-specific errors and reducing the influence of outlier predictions. In this formulation, K denotes the number of models included in each ensemble subset.

$$\hat{P} = \frac{1}{K}\sum_{k=1}^{K} P(y = 1|x, M_k) \quad (3)$$

Performance-Based Weighted Averaging (4 configurations): Unlike simple averaging, this method assigns model-specific weights w_m derived from empirical validation accuracy Acc_m to prioritize the most reliable feature extractors. To ensure a valid probability distribution, weights were normalized as follows:

$$w_m = \frac{Acc_m}{\sum_{j=1}^{M} Acc_j} \quad (4)$$

This ensures that architectures with superior convergence on sMRI biomarkers exert a proportional influence on the final diagnosis.

Third**, Majority (Hard) Voting** was applied to the three-model subsets (A, B, and C) to ensure a non-tie consensus, where the final label C_{final} was determined by the mode of discrete predictions:

$$C_{final} = mode\{arg\,max\,P\,(y|x, M_k)\}_{k=1}^{K} \quad (5)$$

Stacked Generalization (4 configurations): As illustrated in Figure 2, this advanced paradigm treats the probabilistic outputs of base learners as meta-features. A Logistic Regression meta-learner was then trained to exploit inter-model correlations and non-linear decision synergies, effectively compensating for individual architectural limitations by learning the strengths of each model on specific data patterns.

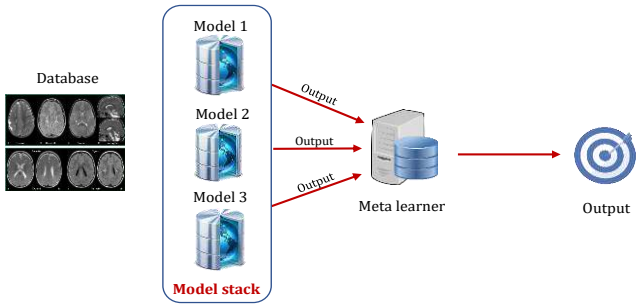


Figure 2: Diagram of stacked generalization (stacking)

To systematically evaluate ensemble performance, all four ensemble types were implemented in four distinct configurations as shown Figure 3 through this multi-strategy ensemble framework, we sought to exploit both output-level integration and meta-learning synergies, ensuring that the ensemble capitalized on the strengths of each architecture while compensating for their individual limitation.

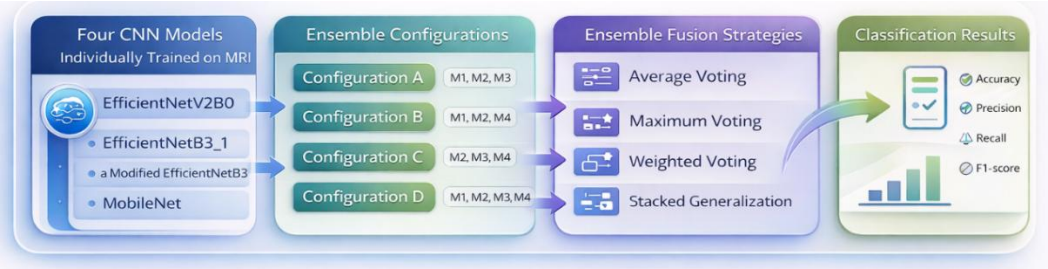


Figure 3: Schematic representation of the Second Experiment.

2.4 Bayesian CNN models and Uncertainty Quantification

To systematically investigate the trade-off between predictive performance, computational efficiency, and diagnostic reliability, three optimized convolutional architectures were employed: two transfer learning backbones—MobileNet and EfficientNetV2-B0—and a custom-developed Lightweight CNN designed from first principles. The pretrained models were initialized with ImageNet weights and structurally adapted for binary ASD classification, while the lightweight architecture was intentionally engineered to minimize parameter complexity (≈2.87 million trainable parameters) and enhance scalability in resource-constrained clinical environments.

For the pretrained Bayesian variants, the feature extraction pipeline was augmented with spatial attention mechanisms to emphasize salient neuroanatomical regions across axial, sagittal, and coronal planes. Extracted representations were compressed via Global Average Pooling (GAP) and subsequently processed through a dual-layer dense classification head (64 units each), incorporating Batch Normalization, dropout regularization with rates of 0.1 and 0.15, and activations functions with an alpha of 0.3 to promote effective learning and reduce overfitting as shown in Figure 5.

In contrast, the custom Lightweight CNN adopted a streamlined hierarchical structure comprising three convolutional blocks, each consisting of convolutional layers with ReLU activation, batch normalization, and max pooling, allowing for effective hierarchical feature extraction and dimensionality reduction. Fully connected dense layers followed, incorporating batch normalization, dropout with rates of 0.1 and 0.2, and ReLU activations to enhance generalization. The model concluded with a sigmoid-activated output neuron to perform binary classification. The complete model architecture is illustrated in Figure 6. Despite its structural simplicity, the lightweight architecture preserved compatibility with the probabilistic modeling framework applied to the pretrained networks.

To overcome the inherent limitations of deterministic convolutional neural networks—particularly their tendency to generate overconfident predictions without calibrated reliability estimates—a Bayesian Deep Learning (BDL) framework was implemented. In Bayesian Convolutional Neural Networks (BCNNs), deterministic weight parameters W are replaced with probability distributions, enabling posterior inference conditioned on the training data $D = (x_i, y_i)$, The posterior distribution is defined via Bayes' theorem (Loredo and Wolpert 2024).

$$P(W|D) = \frac{P(D|W)P(W)}{P(D)} \quad (6)$$

where $P(D|W)$ is the likelihood, $P(W)$ represents the prior distribution, and $P(D)$ is the marginal likelihood. Given the computational intractability of exact posterior estimation in deep architectures with millions of parameters, scalable approximate inference techniques were adopted.

To approximate this intractable posterior, various Bayesian inference techniques exist, primarily Markov Chain Monte Carlo (MCMC), Variational Inference (VI), and Monte Carlo (MC) Dropout (van de Schoot et al. 2021). While MCMC algorithms can theoretically converge to the true exact posterior, they are

computationally prohibitive, highly complex, and practically unfeasible for deep neural networks with millions of parameters (Abdullah A. Abdullah, Hassan, and Mustafa 2022). Consequently, in this research, we deliberately employed VI and MC Dropout as scalable and computationally efficient alternatives to quantify uncertainty.

Specifically, two complementary Bayesian approximation strategies were employed: Variational Inference (VI) and Monte Carlo (MC) Dropout. Variational Inference replaces deterministic dense layers with stochastic DenseVariational layers, modeling weights as Gaussian-distributed random variables $\mathcal{N}(\mu, \sigma^2)$. Training optimizes the Evidence Lower Bound (ELBO), effectively minimizing the Kullback–Leibler (KL) divergence between the approximate posterior $q_\theta(W)$ and the true distribution $P(W|D)$:

$$\mathcal{L}(\theta) = \underbrace{E_{q_\theta(W)}[\log p\,(D|W)]}_{\text{Likelihood Cost}} - \underbrace{\mathrm{KL}\big(q_\theta(W)\,|\,p(W)\big)}_{\text{Complexity Cost}} \qquad (7)$$

This probabilistic formulation enables principled modeling of epistemic uncertainty while regularizing model complexity.

Complementarily, MC Dropout was implemented as a computationally efficient Bayesian approximation by maintaining dropout layers during inference. For each input MRI slice, T stochastic forward passes ($e.g., \boldsymbol{T} = 100$) were performed, yielding a predictive distribution:

$$p(y^*|x^*, D) \approx \frac{1}{T}\sum_{t=1}^{T} p\big(y^*|x^*, \widehat{W}_t\big) \qquad (8)$$

The predictive mean provided the final classification, while the predictive variance served as a quantitative uncertainty metric. This approach offers a practical mechanism for capturing epistemic variability without the prohibitive computational cost associated with Markov Chain Monte Carlo sampling.

Both Bayesian mechanisms (VI and MC Dropout) were independently integrated into the three architectural backbones—MobileNet, EfficientNetV2-B0, and the proposed Lightweight CNN—to ensure a rigorous and architecture-agnostic evaluation of probabilistic inference strategies. In the VI configurations, deterministic dense layers were replaced with DenseVariational layers parameterized by Gaussian priors. In contrast, MC Dropout variants preserved standard dense layers augmented with stochastic dropout (rates 0.1–0.15) retained during inference. A sigmoid-activated output layer generated probabilistic predictions for ASD classification.

The primary objective of this study is to conduct a comprehensive comparative analysis between these Bayesian models and previously developed ensemble strategies. While ensemble learning enhances predictive robustness through output-level integration, it remains fundamentally deterministic. In contrast, Bayesian models provide calibrated uncertainty estimates within a principled probabilistic framework. By systematically evaluating classification accuracy, robustness, computational efficiency, and uncertainty quantification across all configurations, this study aims to determine whether lightweight uncertainty-aware architectures can achieve performance comparable to, or surpassing, ensemble methods while simultaneously enhancing diagnostic transparency and clinical reliability.

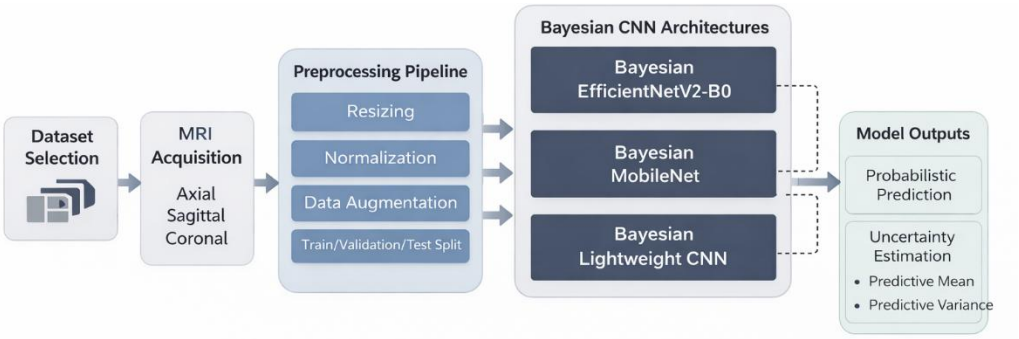


Figure 4: Overview of the Third Experiment: Bayesian Deep Learning Framework

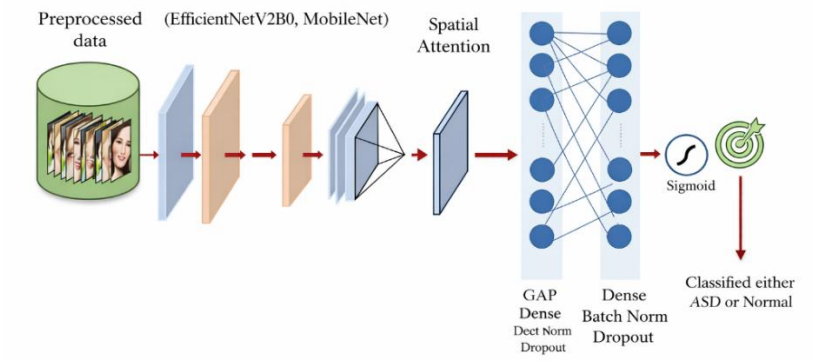


Figure 5: Schematic of the proposed Bayesian pretrained CNN mode

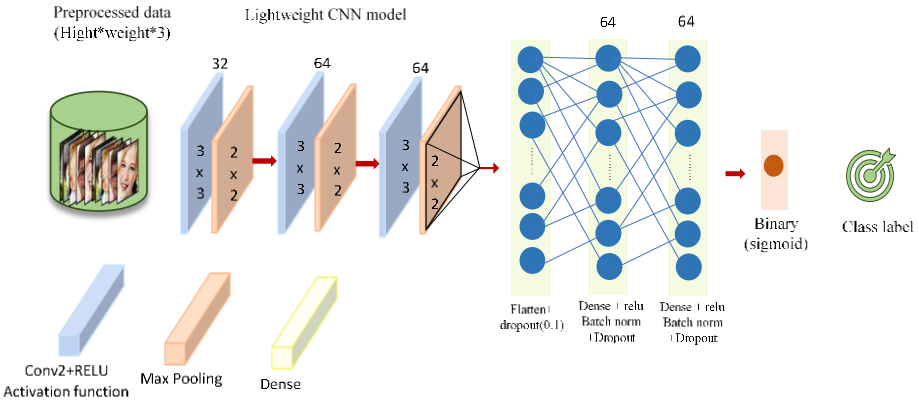


Figure 6: Structure of the proposed lightweight BCNN models for detecting ASD patients.

2.5 Model Training and Testing

A three-phase experimental framework was designed to evaluate predictive performance, generalization, and uncertainty calibration in MRI-based ASD classification. A central component in both experimental setups was the Triangular Cyclical Learning Rate (CLR) scheduler proposed by (Smith 2017), which dynamically varies the learning rate between predefined minimum and maximum bounds. This strategy enhances convergence stability, mitigates the risk of suboptimal local minima, and promotes improved generalization.

For the deterministic baseline, transfer learning was applied using EfficientNetV2B0, both version of EfficientNetB3, and MobileNet pre-trained on ImageNet. The final classification layers were adapted to the target task while preserving pretrained feature representations. All models were trained for 50 epochs using adaptive optimizers. Nadam was employed for the EfficientNet variants with a learning rate range of (1e−6, 1e−3) and batch size of 32, whereas RMSprop was used for MobileNet with a learning rate range of (1e−7, 1e−3) and batch size of 80. Optimization was conducted using categorical cross-entropy.

To model predictive uncertainty, Bayesian CNN frameworks were implemented using Monte Carlo sampling (BCNN-MC) and Variational Inference (BCNN-VI). In BCNN-MC, MobileNet (180 epochs) was optimized with Adamax, while EfficientNetV2B0 (60 epochs) and a lightweight CNN (200 epochs) were trained using Nadam. All models employed a learning rate range of (1e−7, 1e−3), batch size of 32, and binary cross-entropy loss. Predictive posteriors were approximated via multiple stochastic forward passes to estimate both mean predictions and epistemic uncertainty.

For BCNN-VI, MobileNet (160 epochs), EfficientNetV2B0 (60 epochs), and the lightweight CNN (150 epochs) were trained using Nadam under identical learning rate and batch size settings. Optimization minimized the negative log-likelihood objective, integrating expected data likelihood with Kullback–Leibler divergence regularization to approximate the variational posterior distribution. The detailed network configuration is presented Table 3.

All models were evaluated on an unseen hold-out test set. Performance was quantified using accuracy, precision, recall, and F1-score, while probabilistic models were further assessed through uncertainty estimation to ensure rigorous comparison and reliable generalization analysis.

Table 3: Summary of hyperparameters for Individual and probabilistic convolutional models.

The Hyperparameters for Individual Pre-trained Models					
Model	**Epoch**	**Optim.**	**LR Range**	**Batch**	**Loss**
EfficientNetV2B0	50	Nadam	(1-e6, 1-e3)	32	Categorical cross-entropy
EfficientNetB3_1					
EfficientNetB3_1					
MobileNet		RMSprop	(1-e7,1-e3)	80	
The Hyperparameters for BCNN-MC					
Model	**Epoch**	**Optim.**	**LR Range**	**Batch**	**Loss**
MobileNet	180	Adamax	(1-e7, 1-e3)	32	Binary cross-entropy
EfficientNetV2B0	60	Nadam			
lightweight CNN	200				
The Hyperparameters for BCNN-VI					
Model	**Epoch**	**Optim.**	**LR Range**	**Batch**	**Loss**
MobileNet	160	Nadam	(1-e7, 1-e3)	32	Negative log-likelihood
EfficientNetV2B0	60				
lightweight CNN	150				

2.6 Evaluation Metrics

Model performance was evaluated using complementary classification metrics derived from the confusion matrix, including accuracy (ACC), precision (PRE), recall/sensitivity (REC), specificity (SPEC), and F1-score (F1). For deterministic CNNs, metrics were computed directly from the predicted labels, whereas for Bayesian CNNs, predictions were obtained as the mean predictive probability over multiple stochastic forward passes:

$$\bar{p}(y|x) = \left(\frac{1}{T}\right)\sum_{t=1}^{T} p(y|x, w_t) \quad (9)$$

where T is the number of forward passes and w_t are the sampled network weights. Metrics were then calculated using standard formulations:

$$ACC = (TP + TN)/(TP + TN + FP + FN) \quad (10)$$

$$PRE = TP\ /\ (TP + FP) \quad (11)$$

$$REC = TP\ /\ (TP + FN) \quad (12)$$

$$SPEC = TN\ /\ (TN + FP) \quad (13)$$

$$F1 = 2 * (PRE * REC)\ /\ (PRE + REC) \quad (14)$$

3 Result and Discussion

This section presents the experimental results of the proposed models on MRI datasets, covering three main experiments: individual pretrained models, ensemble strategies, and Bayesian approaches. Model performance is evaluated using accuracy, precision, recall, F1-score, and AUC, with Grad-CAM visualizations supporting interpretability.

3.1 Experiment 1: Classification Performance of Individual Pretrained Models

The classification performance of the individual models for MRI datasets is summarized Table 4, using standard evaluation metrics. Four pretrained CNN models (EfficientNetV2B0, EfficientNetB3_1, EfficientNetB3_2, and MobileNet) were evaluated to examine their effectiveness across different input modalities and anatomical views. Evaluated across the axial, sagittal, and coronal anatomical planes, model performance varied among the different views.

In the axial view, EfficientNetB3_2 and EfficientNetV2B0 achieved the highest accuracy (98.88%), demonstrating a strong balance between recall (98.63% and 98.82%) and precision (98.63% and 98.43%), which resulted in a high F1-score of 98.62%. EfficientNetB3_1 also performed competitively in this view, attaining an accuracy of 98.78% and an F1-score of 98.50%. MobileNet achieved an accuracy of 98.41% and an F1-score of 98.05%, indicating consistent and reliable performance across MRI modalities.

Notably, EfficientNetB3_2 achieved the highest performance on the **sagittal MRI**, with an overall accuracy of 98.97%, recall of 98.14%, precision of 99.03%, and an F1-score of 98.70%. EfficientNetV2B0 followed closely, with an accuracy of 98.83% and an F1-score of 98.55%, reflecting robust but slightly lower performance. In contrast, MobileNet showed the least effective results, with an accuracy of 98.42%, recall of 97.29%, and F1-score of 98.04%, suggesting that lightweight architectures may be less capable of capturing the complex structural patterns present in sagittal brain images.

In the coronal view, EfficientNetB3_1 outperformed the other models, achieving the highest accuracy (99.33%), recall (99.07%), precision (99.27%), and F1-score (98.86%), demonstrating strong adaptability to the structural complexity of this orientation. EfficientNetB3_2 also exhibited strong performance, with an accuracy of 99.08% and an F1-score of 98.86%. In contrast, EfficientNetV2B0 showed slightly reduced performance in this view, with an accuracy of 99.03%, precision of 98.83%, and an F1-score of 98.55%, possibly due to misclassifications in regions with anatomical variability. MobileNet continued to demonstrate robust performance, recording an accuracy of 98.63% and an F1-score of 98.31%.

Table 4: Performance of individual pre-trained models used.

Classification Performance of the individual Models for Axial MRI dataset				
Model	ACC. (%)	REC. (%)	PRE. (%)	F1 (%)
EfficientNetV2B0	98.88	98.61	98.63	98.62
EfficientNetB3_1	98.78	98.52	98.49	98.50
EfficientNetB3_2	98.88	98.82	98.43	98.62
MobileNet	98.41	98.07	98.04	98.05
Classification Performance of the individual Models for Sagittal MRI dataset				
Model	ACC. (%)	REC. (%)	PRE. (%)	F1 (%)
EfficientNetV2B0	98.83	98.07	99.04	98.55
EfficientNetB3_1	98.95	98.22	99.19	98.70
EfficientNetB3_2	98.97	98.14	99.03	98.56
MobileNet	98.42	97.29	98.80	98.04
Classification Performance of the individual Models for Coronal MRI dataset				
Model	ACC. (%)	REC. (%)	PRE. (%)	F1 (%)
EfficientNetV2B0	99.03	98.83	98.80	98.81
EfficientNetB3_1	99.33	99.07	99.27	99.17
EfficientNetB3_2	99.08	98.73	99.00	98.86
MobileNet	98.63	98.13	98.48	98.31

To further assess and validate the classification performance of each individual model across the various datasets. As illustrated in Figure 7-9 present the confusion matrices for each model evaluated on the three MRI modalities—axial, sagittal, and coronal, respectively. in the Axial MRI dataset (Figure 7), EfficientNetB3_2 again outperformed the others, with only 52 FP and 39 FN, demonstrating robust feature learning from axial slices. In comparison, EfficientNetV2B0 and B3_1 showed relatively higher FN rates, and MobileNet recorded the highest error rates overall, reducing its diagnostic reliability.

This trend persisted in the sagittal MRI dataset (Figure 8), where EfficientNetB3_2 maintained the lowest misclassification rates (27 FP and 74 FN), indicating consistent performance across views. While B3_1 and V2B0 also performed well, MobileNet again showed the weakest performance with elevated FP and FN rates.

In the coronal MRI dataset (Figure 9), EfficientNetB3_1 emerged as the top performer, with only 24 FP and 31 FN, closely followed by EfficientNetB3_2. Meanwhile, MobileNet continued to display higher misclassification rates, affirming its comparatively limited generalization. to display higher misclassification rates, affirming its comparatively limited generalization.

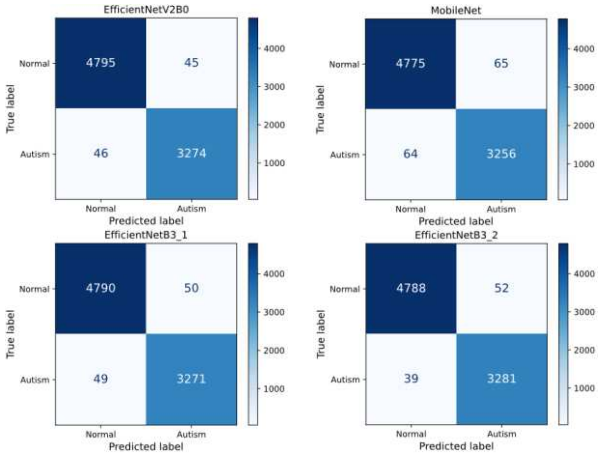


Figure 7: Confusion matrix for each individual pretrained model used for Axial MRI.

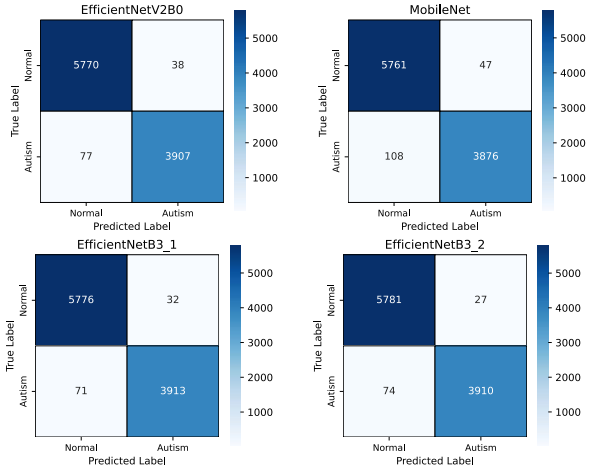


Figure 8*:* Confusion matrix for each individual pretrained model used for Sagittal MRI.

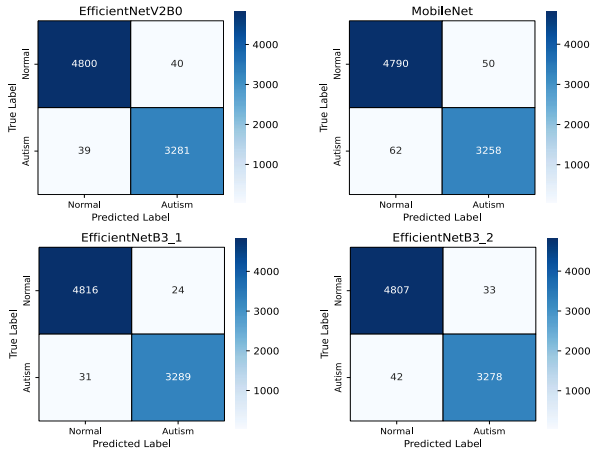


Figure 9: Confusion matrix for each individual pretrained model used for Coronal MRI.

3.2 Experiment 2: Ensemble Model Performance

After evaluating the classification performance of individual pre-trained models in Experiment 1, we extended our investigation in Experiment 2: Ensemble Learning Methods, by employing ensemble learning techniques to combine the strengths of these models as shown in Table 5. The goal was to enhance prediction accuracy, reliability, and robustness, especially in challenging clinical dataset MRI scans.

In the MRI datasets, the stacked ensemble consisting exclusively of EfficientNet models (configuration A) achieved the highest accuracy of 99.20%, surpassing the best individual model, EfficientNetV2B0, which attained 98.88%, marking a relative improvement of +0.32%. Precision, recall, and F1-score similarly improved from 98.62% to 99.02%, demonstrating enhanced diagnostic robustness. The weighted voting ensemble that combined all four models also performed competitively, reaching 99.20% accuracy (+0.32% gain). The average voting ensemble yielded comparable results, with accuracy, recall, precision, and F1-score of 99.19%, 98.86%, 99.15%, and 99.01%, respectively. The majority voting ensemble (configuration A) achieved an accuracy of 98.67%, slightly below the best individual model (98.88%).

Table 5: Evaluation Metrics and Class-wise Accuracy of Ensemble Strategies Across Configurations for Axial MRI.

Output level Ensemble for Axial MRI				
Results of ensemble techniques based on Average voting.				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	99.15	98.86	99.06	98.96
M1, M2, M4	99.11	98.79	99.00	98.90
M2, M3, M4	99.04	98.92	98.73	98.83
M1, M2, M3, M4	99.19	98.86	99.15	99.01
Results of ensemble techniques based on Majority voting.				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	98.67	97.31	99.14	98.22
M1, M2, M4	98.56	97.01	99.44	98.21
M2, M3, M4	98.62	97.14	99.38	98.25
Results of ensemble techniques based on Weighted ensemble.				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	99.15	97.98	99.44	98.70
M1, M2, M4	98.94	98.94	98.94	98.94
M2, M3, M4	98.87	98.87	98.87	98.87
M1, M2, M3, M4	**99.20**	98.79	99.18	98.98
Results of ensemble techniques based on Stacked (Stacking generalization) ensemble.				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	**99.20**	99.60	98.45	99.02
M1, M2, M4	98.95	98.07	99.45	98.76
M2, M3, M4	98.97	98.01	99.39	98.52
M1, M2, M3, M4	99.09	98.31	99.45	98.88

As presented in Table 6 - 7, ensemble learning methods consistently outperformed individual models across both sagittal and coronal MRI datasets, underscoring the effectiveness of model combination strategies.

In the sagittal dataset Table 6, the highest accuracy achieved by individual models was 98.97%, obtained by EfficientNetB3_2. The weighted ensemble and average voting ensemble combining all four models (configuration 4) achieved the highest performance, with an overall accuracy of 99.33% (representing a +0.36% improvement) alongside equally high precision, recall, and F1-score values. This was followed by the stacked ensemble using only EfficientNet models (configuration 1), which achieved 99.14% (+0.19%) with perfectly balanced metrics. Another stacked ensemble (configuration 4) reached 99.18%, further supporting the benefit of this method. Although majority voting yielded more modest results, with accuracies ranging from 98.09% to 98.47%, it remained slightly lower than the best individual model (98.97%).

Table 6*:* Evaluation Metrics of Ensemble Strategies Across Configurations for Sagittal MRI.

Results of ensemble techniques based on Average voting				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	99.24	98.57	99.56	99.06
M1, M2, M4	99.26	98.59	99.59	99.09
M2, M3, M4	99.29	98.64	99.89	99.26
M1, M2, M3, M4	**99.33**	98.72	99.62	99.17
Results of ensemble techniques based on Majority voting				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	98.47	96.33	99.89	98.08
M1, M2, M4	98.09	95.58	99.71	97.60
M2, M3, M4	98.23	95.78	99.87	97.80
Results of ensemble techniques based on Weighted ensemble				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	99.24	98.59	99.59	99.09
M1, M2, M4	99.26	98.59	99.59	99.09
M2, M3, M4	99.29	98.64	99.59	99.11
M1, M2, M3, M4	99.33	98.72	99.62	99.17
Results of ensemble techniques based on Stacked (Stacking generalization) ensemble				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	99.14	99.07	98.82	98.94
M1, M2, M4	98.76	98.14	98.81	98.47
M2, M3, M4	98.79	98.44	98.59	98.51
M1, M2, M3, M4	99.18	99.04	98.95	98.99

A similar trend was observed in the coronal dataset Table 7, where all ensemble techniques—except majority voting—outperformed the individual models. The highest accuracy (99.51%) was achieved by the average voting and weighted ensemble (configuration A), representing a +0.18% improvement over the best individual model (99.33%), with consistently high evaluation metrics. The stacked ensemble (configuration D) and EfficientNet-based ensemble (configuration A) also demonstrated strong performance, achieving accuracies of 99.29% and 99.47%, respectively. Although majority voting yielded lower performance (98.79%), it remained competitive relative to the baseline models.

Table 7: Evaluation Metrics of Ensemble Strategies Across Configurations for Coronal MRI.

Results of ensemble techniques based on Average voting				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	**99.51**	99.43	99.37	99.40
M1, M2, M4	99.39	99.13	99.37	99.25
M2, M3, M4	99.42	99.22	99.37	99.30
M1, M2, M3, M4	99.41	99.17	99.38	99.27
Results of ensemble techniques based on Majority voting				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	98.79	97.34	99.66	98.50
M1, M2, M4	98.69	97.11	99.66	98.37
M2, M3, M4	98.55	96.83	99.60	98.20
Results of ensemble techniques based on Weighted ensemble				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	99.51	99.34	99.37	99.36
M1, M2, M4	99.39	99.13	99.37	99.25
M2, M3, M4	99.39	99.22	99.37	99.30
M1, M2, M3, M4	99.41	99.19	99.37	99.28
Results of ensemble techniques based on Stacked (Stacking generalization) ensemble				
Model	**ACC. (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
M1, M2, M3	99.47	99.38	99.37	99.37
M1, M2, M4	99.23	98.46	99.67	99.06
M2, M3, M4	99.15	98.43	99.48	98.95
M1, M2, M3, M4	99.29	98.64	99.60	99.12

To further improve classification performance, ensemble learning strategies were applied across all imaging modalities, including axial, sagittal, and coronal MRI scans. These ensemble methods consistently outperformed the individual pretrained CNN models, significantly reducing misclassification rates and enhancing predictive accuracy. In the axial MRI dataset, the best performance was observed with the average prediction and weighted ensemble strategies Figure (10-12) rather than majority voting and stacked strategies Figure (11-13) , reflecting stronger generalization and lower error rates compared to individual models.

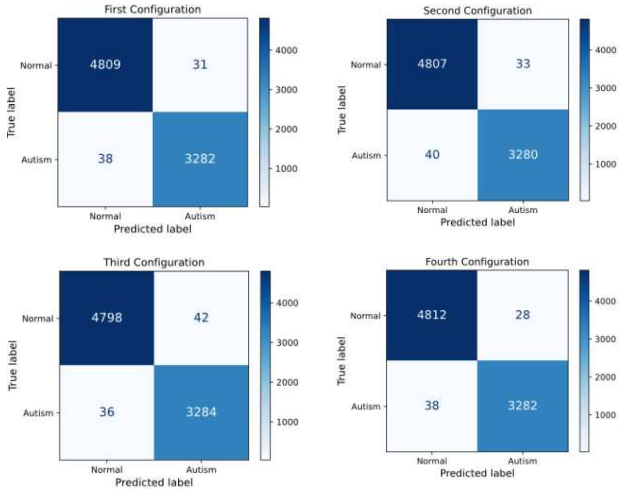


Figure 10*:* Confusion matrices of Average ensemble technique for Axial MRI.

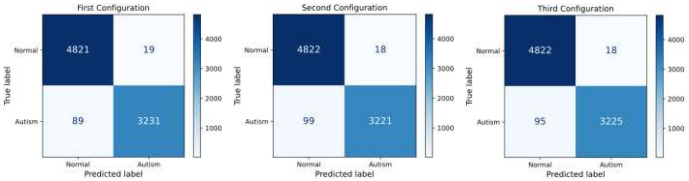


Figure 11: Confusion matrices of majority voting ensemble technique for Axial MRI.

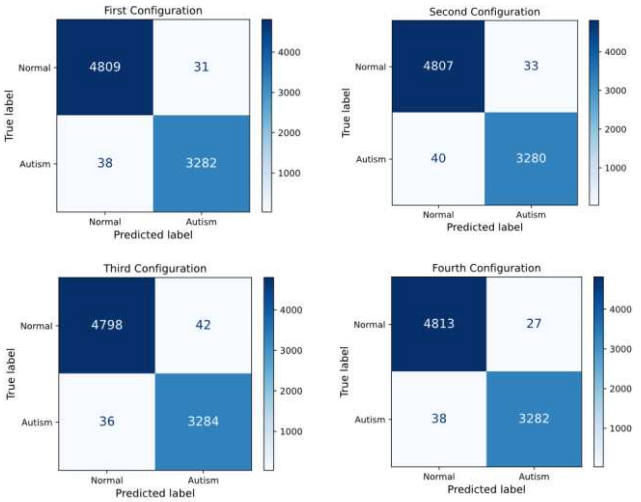


Figure 12: Confusion matrices of Weighted ensemble technique for Axial MRI.

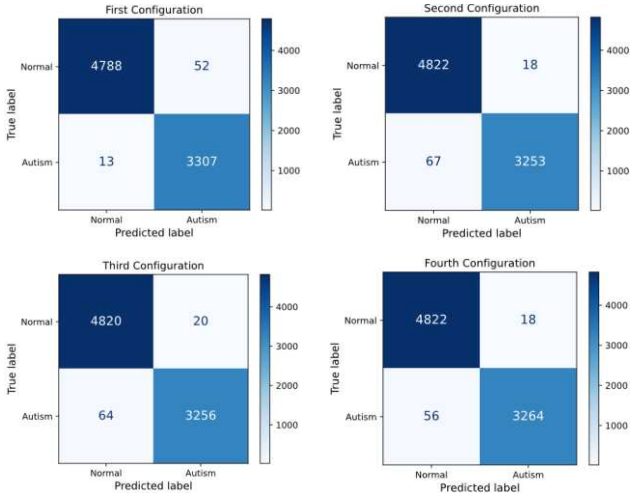


Figure 13: Confusion matrices of Stacked ensemble technique for Axial MRI.

Across sagittal MRI, the average and weighted voting ensembles consistently achieved the best results (Figure 14 and 16), while for coronal MRI, with weighted and average prediction both showing robust performance (Figure 20 and 18)

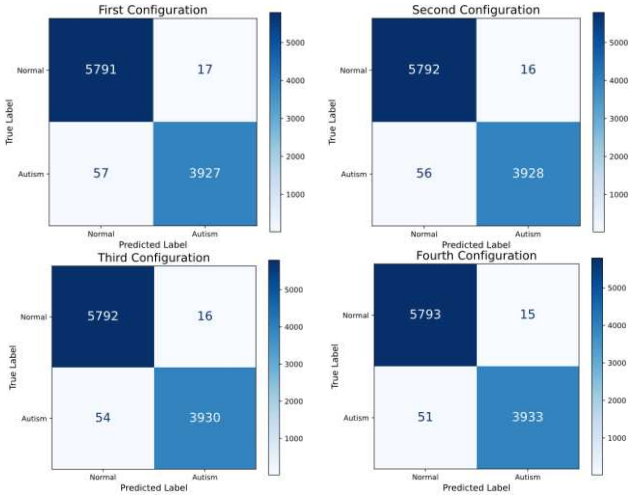


Figure 14: Confusion matrices of Average ensemble technique for Sagittal MRI.

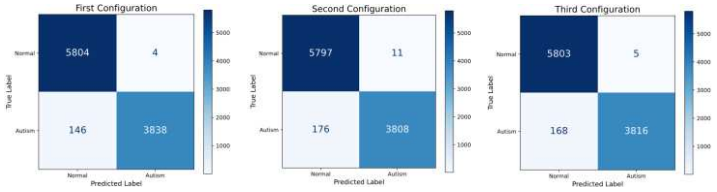


Figure 15: Confusion matrices of Majority voting ensemble technique for Sagittal MRI.

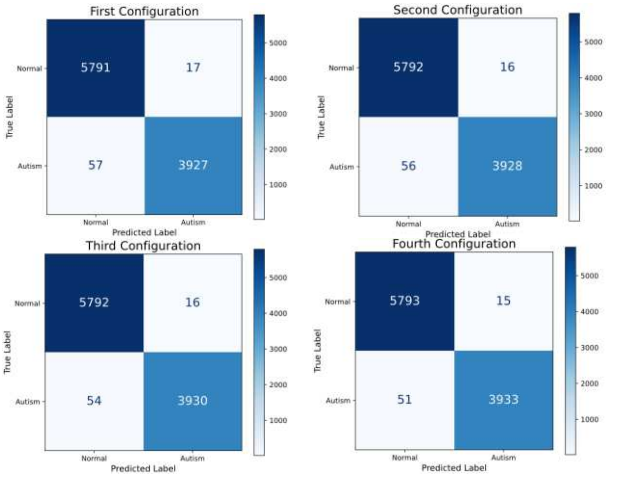


Figure 16: Confusion matrices of Weighted ensemble technique for Sagittal MRI.

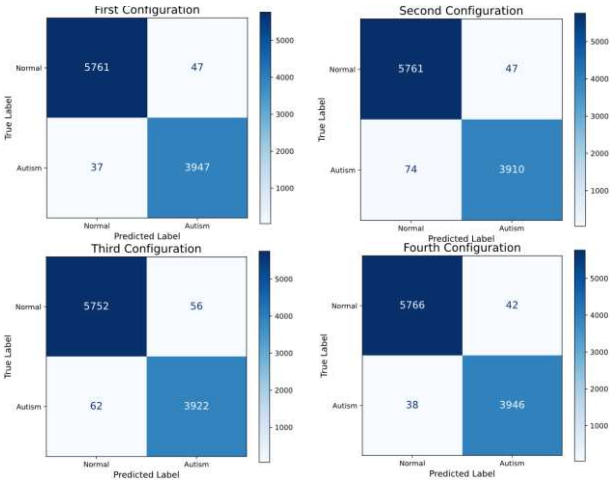


Figure 17: Confusion matrices of Stacked ensemble technique for Sagittal MRI.

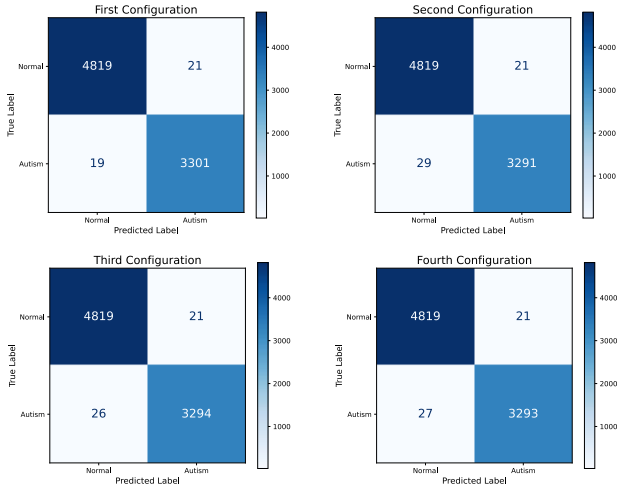


Figure 18: Confusion matrices of Average ensemble technique for Coronal MRI

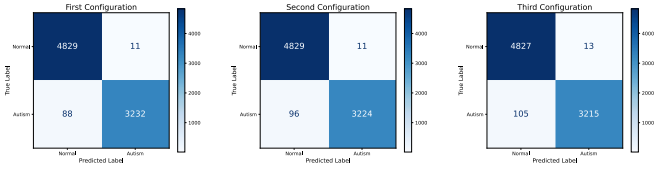


Figure 19*:* Confusion matrices of Majority voting ensemble technique for Coronal MRI.

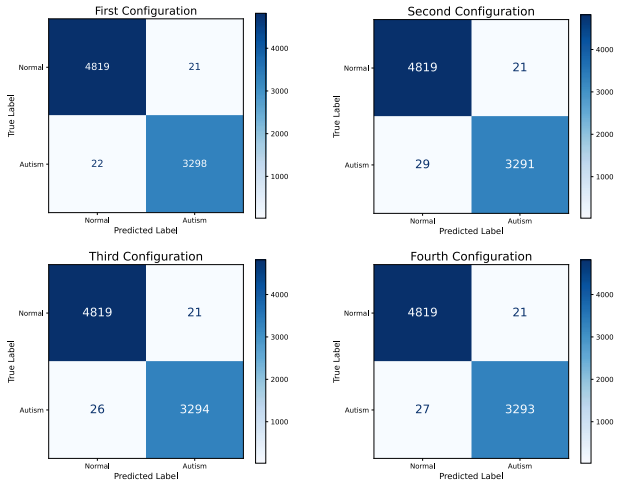


Figure 20: Confusion matrices of Weighted ensemble technique for Coronal MRI.

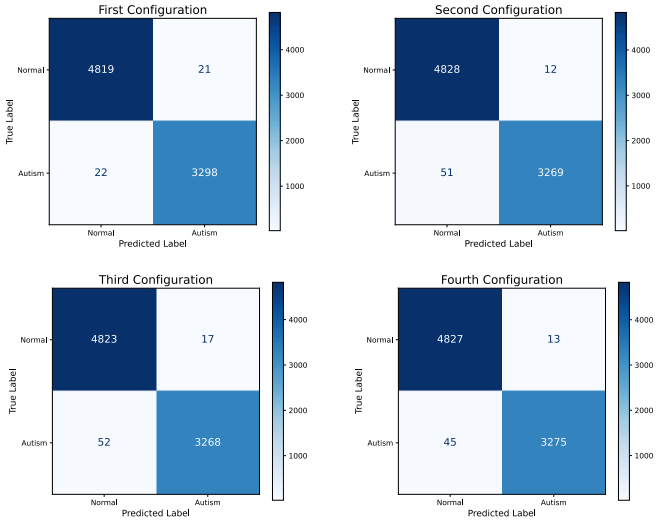


Figure 21: Confusion matrices of Stacked ensemble technique for Coronal MRI.

3.3 Experiment 3: Performance of Proposed Bayesian CNN Models

Following the comprehensive evaluation of ensemble-based methods, the investigation advanced toward the development of lightweight attention-enhanced models integrated with Bayesian MC-Dropout and VI, two prominent techniques for uncertainty quantification in DL. These models were designed to sustain high diagnostic accuracy while simultaneously reducing computational complexity and enhancing interpretability.

In this context, the effect of incorporating uncertainty estimation into the proposed model configurations was investigated using the MRI dataset. Specifically, MC-Dropout was employed to quantify predictive uncertainty by performing multiple stochastic forward passes during inference. Unlike conventional deterministic models that yield a single-point prediction, MC-Dropout retains dropout layers at test time and executes, for example, 100 forward passes per image, resulting in a distribution of predictions. The mean of this distribution was used to derive the final classification, while the standard deviation represented the model's confidence. This process enabled the quantification of epistemic uncertainty, which arises from uncertainty in the model parameters and is particularly valuable for assessing the reliability of predictions in high-stakes clinical applications. As demonstrated in table 5 to 7, the most effective ensemble configurations, particularly the weighted and stacked ensembles combining M_1, M_2, M_3, and M_4, achieved classification accuracies of 99.20% (Axial), 99.33% (Sagittal), and 99.51% (Coronal).

In comparison, Bayesian MC-Dropout-based lightweight models demonstrated highly competitive performance, as summarized in Table 8 MobileNet achieved accuracies of 98.54%, 98.94%, and 98.77%; EfficientNetV2B0 attained 98.87%, 99.51%, and 99.17%; and the lightweight CNN achieved 98.05%, 99.30%, and 99.19% across the respective MRI views. Although ensemble models yielded marginal improvements, typically within a 0.13%–0.32% range, the lightweight MC-Dropout models—particularly EfficientNetV2B0—produced comparable or even superior results. Moreover, unlike conventional ensembles, these Bayesian models inherently quantify predictive uncertainty, offering both competitive accuracy and enhanced reliability for clinical decision support. Notably, EfficientNetV2B0 surpassed the best ensemble model in the sagittal view by 0.18%, further underscoring its robustness and practical advantage.

In contrast, VI-based lightweight models, summarized in Table 8, exhibited noticeably weaker performance. While MobileNet and EfficientNetV2B0 still achieved acceptable accuracies (e.g., 98.90% and 98.44% in the Axial view, respectively), their results were less consistent across views and generally inferior to MC-Dropout-based models. Most significantly, the lightweight CNN trained with VI performed poorly, with accuracies declining to 84.10% (Axial), 81.17% (Sagittal), and

85.07% (Coronal), marking the lowest performance across all experimental configurations. This stark contrast indicates that although VI offers a theoretically principled approach for capturing epistemic uncertainty, its practical effectiveness in lightweight attention-enhanced architectures is limited.

Table 8: Performance of Bayesian Models with MC-Dropout and VI Trained on MRI Views.

MC Dropout-Based Model					
Trained on Axial MRI Views					
Model	**Acc (%)**	**Std (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
MobileNet	98.54	0.087	97.83	98.72	98.27
EfficientNetV2B0	**98.87**	0.032	98.25	99.08	98.69
Lightweight CNN	98.05	0.032	98.91	98. 86	98.88
Trained on Sagittal MRI Views					
Model	**Acc (%)**	**Std (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
MobileNet	98.94	0.052	97.89	99.49	98.69
EfficientNetV2B0	**99.51**	0.018	99.30	99.80	99.55
Light weight CNN	99.30	0.027	99.20	99.90	99.55
Trained on Coronal MRI Views					
Model	**Acc (%)**	**Std (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
MobileNet	98.77	0.064	98.31	98.60	98.45
EfficientNetV2B0	99.17	0.013	98.92	99.09	99.00
Light weight CNN	**99.19**	0.044	99.16	98.86	99.01
VI_Based Model					
Trained on Axial MRI Views					
Model	**Acc (%)**	**Std (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
MobileNet	98.90	2.107	98.67	98.73	98.70
EfficientNetV2B0	98.44	2.56	97.80	98.54	98.17
Light weight CNN	84.10	40.30	84.63	79.37	81.91
Trained on Sagittal MRI Views					
Model	**Acc (%)**	**Std (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
MobileNet	98.51	1.50	98.64	98.68	98.66
EfficientNetV2B0	99.04	2.22	99.04	99.04	99.04
Light weight CNN	81.17	57.48	76.40	76.70	76.55
Trained on Coronal MRI Views					
Model	**Acc (%)**	**Std (%)**	**REC. (%)**	**PRE. (%)**	**F1 (%)**
MobileNet	98.68	1.30	98.43	98.19	98.30
EfficientNetV2B0	98.63	1.81	98.92	99.03	98.97
Light weight CNN	85.07	40.94	76.26	76.40	76.33

The confusion matrices presented in (Figures 22 to 24) demonstrate that Bayesian MC Dropout models consistently achieve superior classification performance compared to their VI counterparts across all anatomical MRI views. Specifically, MobileNet and EfficientNetV2B0 architectures with MC Dropout exhibit balanced and accurate predictions, characterized by minimal misclassifications, whereas the VI variants show increased error rates and diminished generalization ability. This disparity is most pronounced in the Lightweight CNN, where the MC Dropout model maintains reasonable accuracy, but the VI-based model suffers from significant misclassification and instability. These findings align with earlier results in accuracy and F1-score metrics, reinforcing that MC Dropout offers more reliable and robust uncertainty estimation than VI, especially in lightweight models applied to clinical autism diagnosis.

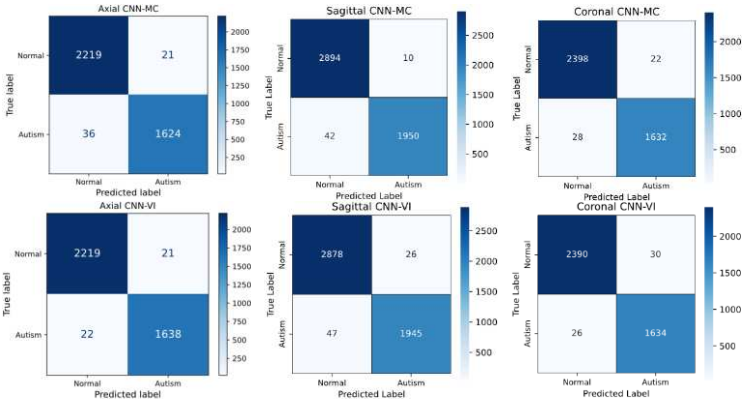


Figure 22*:* Confusion matrices of MobileNet model using Bayesian MC-Dropout and VI across anatomic of MRI.

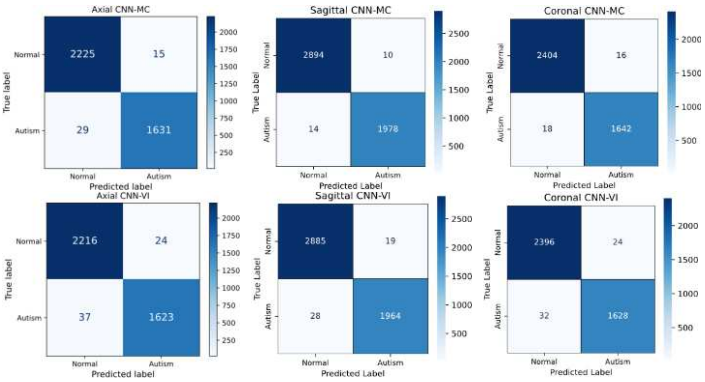


Figure 23: Confusion matrices of EfficientNetV2B0 model using Bayesian MC-Dropout and VI across anatomic of MRI.

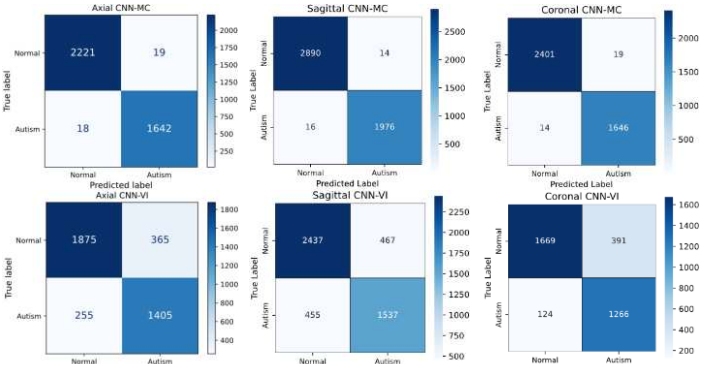


Figure 24: Confusion matrices of Lightweight CNN model using Bayesian MC-Dropout and VI across anatomic of MRI.

3.4 Uncertainty Quantification Analysis

To gain a deeper understanding of predictive uncertainty, we analyzed the distribution of predicted probabilities for sixteen representative test samples by performing 1,000 stochastic forward passes through Bayesian deep learning models implemented using both MC-Dropout and VI. In this framework, epistemic uncertainty is quantified as the standard deviation of predicted probabilities across these stochastic passes, where a low standard deviation corresponds to high model confidence due to consistent predictions, and a high standard deviation indicates greater model uncertainty caused by variability in the posterior samples. Figure 31 present several representative samples, displaying the mean probability maps alongside the corresponding standard deviation (std) maps, which together characterize the model's confidence and uncertainty. It is important to note that a sigmoid activation function was used, such that mean values close to 0 correspond to the Normal class, while values close to 1 indicate the ASD class.

As illustrated in Figure 31, the model exhibits varying degrees of confidence across samples. For instance, Sample B, which correctly classified as ASD, demonstrates a high mean prediction of 0.99 paired

with a low standard deviation of 0.02, signifying a very confident and reliable classification. By contrast, Sample D, which classified as Normal, has a low mean of 0.00 and a relatively low standard deviation of 0.03, indicating the model's strong leaning towards the Normal class with limited uncertainty. However, Sample E reveals more complexity, showing a moderate mean of 0.38 alongside a high standard deviation of 0.26. While the mean suggests a Normal classification, the substantial uncertainty and the true label being Autism highlight the model's difficulty in making a confident assignment, particularly near the decision boundary.

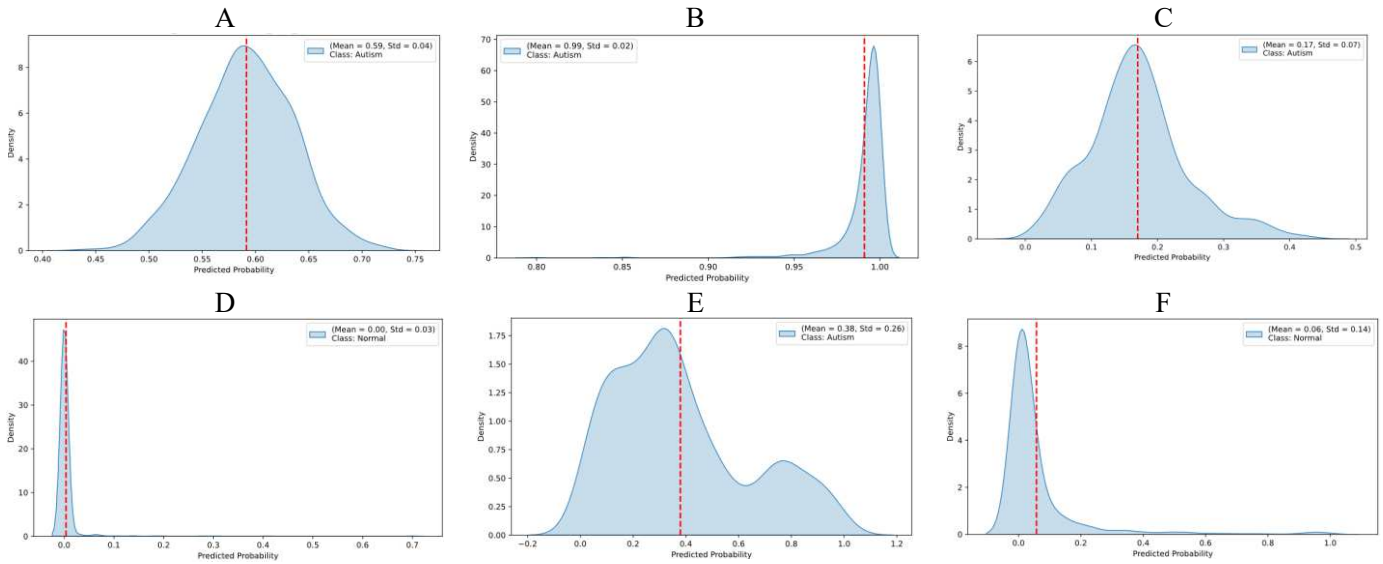


Figure 25: Posterior distributions of several predicted examples with their associated uncertainties from the Bayesian MC-Dropout model, applied to the MRI dataset.

3.5 Analysis of Gradient-weighted Class Activation Maps

Gradient-weighted Class Activation Maps (Grad-CAMs) generated from the CNN model to visualize the brain regions most influential in distinguishing between autistic and non-autistic subjects. Figure 32 corresponds to the autism group, while Figure 33 represents the normal (non-autistic) group. Each figure presents Grad-CAMs across three anatomical planes—axial, sagittal, and coronal—providing a comprehensive spatial interpretation of the model's focus during classification.

In Figure 32, the Grad-CAMs for autistic subjects highlight distinct activation regions primarily concentrated in the frontal, temporal, and parietal lobes, as well as subcortical areas such as the amygdala and thalamus. These regions are known to be associated with social cognition, emotional regulation, and communication processing, which are often altered in Autism Spectrum Disorder (ASD). The consistent activation patterns across multiple views (A–E) suggest that the CNN model has learned to identify stable morphological or functional signatures characteristic of autism.

In contrast, Figure 33 depicts the Grad-CAMs for normal subjects, where the highlighted areas are more evenly distributed and less intense. The model's attention appears to focus on structurally typical regions of the cortex, particularly within the occipital and parietal lobes, indicating typical neural patterns associated with normal development. The difference in the intensity and localization of highlighted regions between the two figures suggests that the CNN effectively distinguishes autism-related abnormalities from typical brain patterns.

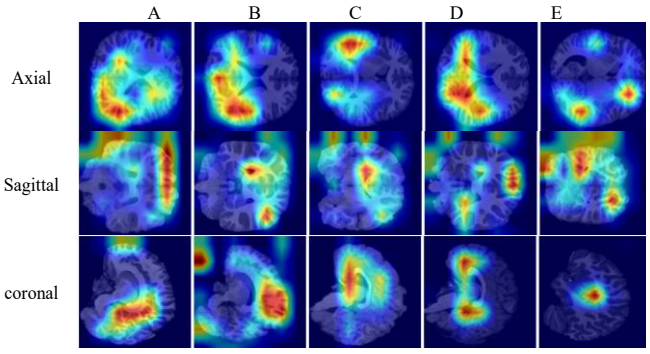


Figure 26: Grad CAMs for CNN model for autism subject across axial, sagittal, and coronal views, highlighting regions contributing most to classification.

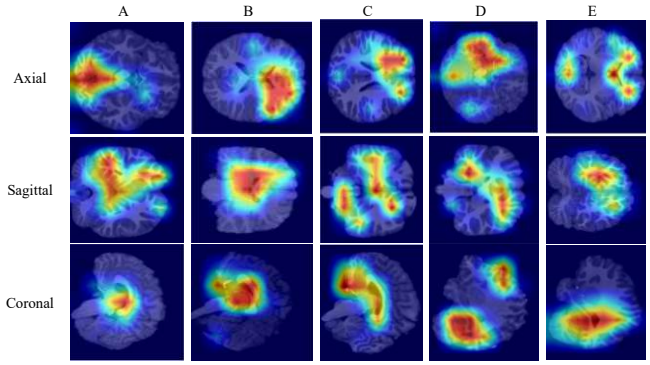


Figure 27*:* Grad-CAMs of the CNN model for a normal subject across axial, sagittal, and coronal views, highlighting regions contributing most to classification.

4 Comparison with State-of-the-art Methods for MRI Dataset

To validate the effectiveness and generalizability of our proposed models, we conducted a comparative analysis against state-of-the-art methods for MRI-based ASD classification, as well as between our lightweight model and the ensemble framework. This comparison demonstrates clear improvements in accuracy, robustness, and clinical applicability.

4.1 Comparison with State-of-the-Art Methods for MRI Dataset

To evaluate the effectiveness of our proposed MRI-based classification framework, we compared it with several state-of-the-art methods reported in recent literature. The ensemble model and BCNN model selected for this comparison are the one that demonstrated the best performance across all MRI views in our experiments. The results, summarized in Table 9 include key performance metrics such as accuracy, precision, recall, and F1-score, which is essential measures of model effectiveness and clinical applicability. As can be seen from results in Table 4-8 the proposed ensemble model achieved accuracies of 99.20%, 99.33%, and 99.51% for the axial, sagittal, and coronal views, respectively, placing it among the highest-performing models in the literature. It performs comparably to highly parameterized CNN architectures like VGG16, which reported 99.83% accuracy on the ABIDE dataset (Jain et al. 2023), and CNN models trained on site-specific datasets such as NYU (Feng and Xu 2023) and UCLA (Darweesh et al. 2022) that achieved accuracies of 99.39% and 99.80%, respectively. However, unlike these models, which rely on complex architectures with large parameter counts (e.g., VGG16's ~138 million parameters), our ensemble approach maintains high accuracy with significantly lower computational demands, enhancing its suitability for clinical deployment.

In addition to the ensemble approach, we incorporated lightweight BCNN models utilizing VI and MC-Dropout to quantify predictive uncertainty. The VI-based models achieved accuracies of 98.90%, 99.04%, and 98.68% across the axial, sagittal, and coronal views, respectively, while the MC Dropout models attained accuracies of 98.87%, 99.51%, and 99.19% for the corresponding views. This uncertainty estimation capability provides vital interpretability and clinical trustworthiness, which remains largely absent in previous studies. In contrast, other recent works such as (Mostafa et al. 2023) with Inception-V4 and (Parui et al. 2023) reported lower accuracies between 84.04% and 84.79%, potentially due to conventional architectures and limited preprocessing. Saponaro et al. (Saponaro et al. 2024) faced challenges with fusion-based approaches, reporting low accuracy (85.00%). Although Prasad et al. (Prasad, G. V. Sriramakrishnan, and Diana Jeba Jingle 2023) achieved 95.7% accuracy with a hybrid optimization method, their model neither provides UQ nor demonstrates scalability.

In summary, our ensemble-based framework achieves state-of-the-art performance by balancing accuracy, computational efficiency, and uncertainty estimation. This combination ensures enhanced generalizability and robustness compared to existing methods, making it a more practical and reliable solution for MRI-based autism spectrum disorder classification in diverse clinical settings.

4.2 Comparison between Lightweight BCNNs and Ensemble Methods

The comparison between Bayesian lightweight models and ensemble approaches highlights their complementary strengths in MRI-based ASD classification. Architecturally, ensemble methods achieve robustness by aggregating multiple independently trained CNNs—such as EfficientNetV2B0, EfficientNetB3 variants, and MobileNet—allowing each model to contribute distinct feature representations. This structural redundancy enhances predictive performance by reducing individual bias and variance, but comes at the cost of increased computational overhead, memory usage, and inference time. In contrast, lightweight Bayesian Convolutional Neural Networks (BCNNs) achieve robustness within a single, compact architecture by incorporating probabilistic modeling directly into the network parameters. Uncertainty is captured through Monte Carlo (MC) Dropout or Variational Inference (VI), enabling posterior weight approximation and stochastic forward passes without requiring multiple models. This design results in a significantly smaller parameter footprint (~2.87M), making BCNNs highly suitable for deployment in resource-constrained clinical environments.

From a performance standpoint, ensemble strategies—particularly weighted voting and stacked generalization—attained the highest accuracies across axial, sagittal, and coronal views (99.07%, 99.33%, and 99.51%, respectively), demonstrating the efficacy of combining diverse architectures to mitigate inter-subject and inter-site variability. Meanwhile, Bayesian lightweight models achieved comparably strong results (axial: 98.93%; sagittal: 99.51%; coronal: 99.19%) while providing calibrated uncertainty estimates. This uncertainty quantification enables clinicians to differentiate confident predictions from ambiguous or out-of-distribution cases, enhancing transparency and clinical reliability. Although VI exhibited slightly lower performance compared to MC Dropout, the overall Bayesian framework effectively balances diagnostic accuracy, computational efficiency, and uncertainty awareness—three critical factors for practical medical imaging applications.

Taken together, ensemble methods are optimal for maximizing classification accuracy and robustness, whereas lightweight BCNNs excel in clinical applicability by integrating efficiency, transparency, and principled uncertainty estimation. This dual perspective underscores the complementary nature of the two approaches and provides a foundation for developing robust, interpretable, and deployable ASD diagnostic systems.

Table 9: Comparison of the proposed method with previous studies on the MRI dataset

Author	Sample size	Model	ACC. (%)	REC. (%)	PRE. (%)	F1 (%)
al. (Darweesh et al. 2022)	141 UCLA	CNN model	99.80	-	-	-
(Feng and Xu 2023)	126 NYU	CNN model	99.39	99.85	98.80	99.32
(Herath et al. 2021)	138 KKI	Inception V3	98.35	98.60	99.27	98.89
(Mostafa et al. 2023)	ABIDE II, I	ResNet50, VGG-16, Inception-V4	84.04, 83.22	-	-	-

(Jain et al. 2023)	ABIDE	VGG16	99.83	99.85	99.85	-
(Saponaro et al. 2024)	Preprocessed (CPAC)	joint fusion-DL	85.00	-	-	-
(Prasad, G. V. Sriramakrishnan, et al. 2023)	ABIDE	hybrid sewing optimization	95.70	-	-	-
(Parui et al. 2023)	Preprocessed (CPAC)	LERT Based-CNN	84.79	89.63	78.96	-
Proposed Model	NYU (27) SU (42), EMC (52), KKI (211), UM (28), GU (48) Total =408	Ensemble	Axial=99.20 Sagittal=99.33 Coronal =99.51	98.45, 99.62 99.37	99.60 98.72 99.43	99.02 99.17 99.40
		Lightweight-BCNN(MCD)	Axial=98.87 Sagittal=99.51 Coronal =99.19	99.08, 99.80 98.86	98.25 99.30 99.16	98.69 99.55 99.01
		Lightweight-BCNN(VI)	Axial=98.90 Sagittal=99.04 Coronal =98.68	98.73 99.04 98.19	98.67 99.04 98.43	98.70 99.04 98.30

7. Conclusion and Future Works

In conclusion, this study demonstrates that both ensemble learning and Bayesian CNN methodologies enhance the reliability and robustness of ASD classification using MRI data from the ABIDE II repository. The research was conducted in three phases: first, fine-tuning multiple pretrained CNN models to establish strong baseline performance; second, evaluating ensemble strategies—including stacked and weighted averaging—which delivered the strongest and most stable predictive performance; and third, developing lightweight Bayesian CNNs (BCNNs) that achieved competitive results compared with ensemble models while providing uncertainty estimation critical for clinical decision-making in complex or ambiguous cases. This comparative analysis highlights the complementary benefits of the two approaches: ensembles maximize predictive strength and stability, whereas BCNNs offer efficient, uncertainty-aware predictions that enhance interpretability and clinical trust. Despite these advantages, ensemble methods involve higher computational and memory demands, and Bayesian approaches may face scalability challenges with high-dimensional MRI data. Future research should focus on developing more efficient BCNN architectures, advanced data augmentation and sampling strategies to mitigate class imbalance, and privacy-preserving federated learning frameworks to support robust, clinically deployable ASD diagnostic systems. Collectively, these findings demonstrate that integrating ensemble and Bayesian strategies across a multi-phase experimental framework provides a balanced, interpretable, and reliable approach for AI-assisted ASD diagnosis.

Reference

Abdullah, Abdullah A, Masoud M. Hassan, and Yaseen T. Mustafa. 2022. "A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges." *IEEe Access* 10:36538–62.

Abdullah, Abdullah A., Masoud M. Hassan, and Yaseen T. Mustafa. 2022. "A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges." *IEEE Access* 10:36538–62. doi: 10.1109/ACCESS.2022.3163384.

Anon. 2023. "Keras Applications." *Keras* Keras Applications. Retrieved June 17, 2025 (https://keras.io/api/applications/).

Darweesh, Asmaa N., Nancy M. Salem, and Walid Al-Atabany. 2022. "Classification of Autism Spectrum Disorder Using Convolutional Neural Networks from Neuroimaging Data." *NILES 2022 - 4th Novel Intelligent and Leading Emerging Sciences Conference, Proceedings* 276–79. doi: 10.1109/NILES56402.2022.9942419.

Feng, Min, and Juncai Xu. 2023. "Detection of ASD Children through Deep-Learning Application of FMRI." *Children* 10(10):1–17. doi: 10.3390/children10101654.

Ganaie, M. A., Minghui Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan. 2022. "Ensemble Deep Learning: A Review." *Engineering Applications of Artificial Intelligence* 115:105151. doi: https://doi.org/10.1016/j.engappai.2022.105151.

Goan, Ethan, and Clinton Fookes. 2020. "Bayesian Neural Networks: An Introduction and Survey BT - Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018." Pp. 45–87 in, edited by K. L. Mengersen, P. Pudlo, and C. P. Robert. Cham: Springer International Publishing.

Herath, Lakmini, Dulani Meedeniya, M. A. J. C. Marasingha, and Vajira Weerasinghe. 2021. "Autism Spectrum Disorder Diagnosis Support Model Using Inception V3." *Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2021* (December):1–7. doi: 10.1109/SCSE53661.2021.9568314.

Jain, Sweta, Hrudaya Kumar Tripathy, Saurav Mallik, Hong Qin, Yara Shaalan, and Khaled Shaalan. 2023. "Autism Detection of MRI Brain Images Using Hybrid Deep CNN With DM-Resnet Classifier." *IEEE Access* 11(October):117741–51. doi: 10.1109/ACCESS.2023.3325701.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521(7553):436–44. doi: 10.1038/nature14539.

Lin, Xu, Gui Bing Chen, Ke Yi Li, and Ya Juan Xu. 2025. "Causal Relationship between IgA Nephropathy and Common Neuropsychiatric Disorders: A Two-Sample Mendelian Randomization Analysis." *Journal of Affective Disorders* 369(October 2024):782–88. doi: 10.1016/j.jad.2024.10.012.

Loredo, Thomas J., and Robert L. Wolpert. 2024. "Bayesian Inference: More than Bayes's Theorem." *Frontiers in Astronomy and Space Sciences* 11(October):1–19. doi: 10.3389/fspas.2024.1326926.

Mostafa, Sadab, Zihadul Karim, and Muhammad Iqbal Hossain. 2023. "Autism Detection Based on MRI Images Using Deep Learning." *2023 26th International Conference on Computer and Information Technology, ICCIT 2023* (January). doi: 10.1109/ICCIT60459.2023.10441006.

Panwar, Harsh, P. K. Gupta, Mohammad Khubeb, and Ruben Morales-menendez. 2020. "Since January 2020 Elsevier Has Created a COVID-19 Resource Centre with Free Information in English and Mandarin on the Novel Coronavirus COVID- 19 . The COVID-19 Resource Centre Is Hosted on Elsevier Connect , the Company ' s Public News and Information ." (January).

Parui, Sricheta, Debasis Samanta, Nishant Chakravorty, Uttam Ghosh, and Joel J. P. C. Rodrigues. 2023. "Artificial Intelligence and Sensor-Based Autism Spectrum Disorder Diagnosis Using Brain Connectivity Analysis." *Computers and Electrical Engineering* 108:1–20. doi: 10.1016/j.compeleceng.2023.108720.

Prasad, Vadamodula, G. V. Sriramakrishnan, and I. Diana Jeba Jingle. 2023. "Autism Spectrum Disorder Detection Using Brain MRI Image Enabled Deep Learning with Hybrid Sewing Training Optimization." *Signal, Image and Video Processing* 17(8):4001–8. doi: 10.1007/s11760-023-02630-y.

Prasad, Vadamodula, G. V Sriramakrishnan, and I. Diana Jeba Jingle. 2023. "Autism Spectrum Disorder Detection Using Brain MRI Image Enabled Deep Learning with Hybrid Sewing Training Optimization." *Signal, Image and Video Processing* 17(8):4001–8.

Salehi, Ahmad Waleed, Shakir Khan, Gaurav Gupta, Bayan Ibrahimm Alabduallah, and Abrar Almjally. 2023. "Cnn1.Pdf."

Saponaro, Sara, Francesca Lizzi, Giacomo Serra, Francesca Mainas, Piernicola Oliva, Alessia Giuliano, Sara Calderoni, and Alessandra Retico. 2024. "Deep Learning Based Joint Fusion Approach to Exploit Anatomical and Functional Brain Information in Autism Spectrum Disorders." *Brain Informatics* 11(1). doi: 10.1186/s40708-023-00217-4.

van de Schoot, Rens, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G. Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, and Christopher Yau. 2021. "Bayesian Statistics and Modelling." *Nature Reviews Methods Primers* 1(1):1. doi: 10.1038/s43586-020-00001-2.

Smith, Leslie N. 2017. "Cyclical Learning Rates for Training Neural Networks." *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017* (April):464–72. doi: 10.1109/WACV.2017.58.

Wang, Hao, and Dit Yan Yeung. 2021. "A Survey on Bayesian Deep Learning." *ACM Computing Surveys* 53(5). doi: 10.1145/3409383.