Uncertainty in Deep Learning for EEG under Dataset Shifts

Mats Tveter [1,2], Thomas Tveitstøl [1,2], Christoffer Hatlestad-Hall [1], Hugo L. Hammer [3,4,*], and Ira R. J. Hebold Haraldsen [1,*]

[1] Department of Neurology, Oslo University Hospital, Oslo, Norway
[2] Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[3] Department of Holistic Systems, SimulaMet, Oslo, Norway
[4] Department of Computer Science, Oslo Metropolitan University, Oslo, Norway
∗ Shared last authorship

E-mail: mhtveter@uio.no

May 2025

Abstract.
Objective: As artificial intelligence (AI) is increasingly integrated into medical diagnostics, it is essential that predictive models provide not only accurate outputs but also reliable estimates of uncertainty. In clinical applications, where decisions have significant consequences, understanding the confidence behind each prediction is as critical as the prediction itself. Uncertainty modelling plays a key role in improving trust, guiding decision-making, and identifying unreliable outputs, particularly under dataset shift or in out-of-distribution settings. The primary aim of uncertainty metrics is to align model confidence closely with actual predictive performance, ensuring confidence estimates dynamically adjust to reflect increasing errors or decreasing reliability of predictions. This study investigates how different ensemble learning strategies affect both performance and uncertainty estimation in a clinically relevant task: classifying Normal, Mild Cognitive Impairment, and Dementia from electroencephalography (EEG) data.

Approach: We evaluated the performance and uncertainty of ensemble methods and Monte Carlo dropout on a large EEG dataset. The models were assessed in three settings: (1) in-distribution performance on a held-out test set, (2) generalisation to three out-of-distribution datasets, and (3) performance under gradual, EEG-specific dataset shifts simulating noise, drift, and frequency perturbation.

Main results: Ensembles consisting of multiple independently trained models, such as deep ensembles, consistently achieved higher performance in both the in-distribution test set and the out-of-distribution datasets. These models also produced more informative and responsive uncertainty estimates under various types of EEG dataset shifts.

Significance: These results highlight the benefits of ensemble diversity and independent training to build robust and uncertainty-aware EEG classification models. The findings are particularly relevant for clinical applications, where reliability under distribution shift and transparent uncertainty are essential for safe deployment.

Keywords: AI, Uncertainty, DL, dataset shifts, EEG

1. Introduction

Recent advances in artificial intelligence (AI) have significantly enhanced its capabilities, establishing it as a critical component in various scientific and industrial applications. One of the fields where this technology has the greatest potential for impact is medical diagnostics and risk assessment, with predictive algorithms, research methodologies, and data analysis [1, 2]. Deep learning (DL), a subfield of AI [3], is notable for the depth and complexity of its models, which enables them to learn directly from raw data. These models often contain millions to billions of parameters and have achieved state-of-the-art performance across a wide range of tasks. In low-stakes or non-critical settings, such complexity is generally unproblematic, as occasional prediction errors are unlikely to have serious consequences. However, in clinical medicine, where predictions can have significant consequences for individuals, this complexity introduces critical challenges and safety concerns [4]. The potential for inaccurate predictions in high-stakes settings requires extensive evaluation and validation to ensure safe and effective integration of these models into clinical practice [5]. Two main strategies are commonly used to address the challenges of complexity and trust in AI; explainability and uncertainty modelling. Explainability methods aim to improve interpretability by clarifying the decision-making processes of AI models [6, 7], but do not assess the reliability of the decision [8] and are vulnerable to overconfident predictions and inputs specifically designed to mislead the model, known as adversarial attacks [9]. On the other hand, uncertainty modelling aims to inform users about the confidence of predictions, thereby enhancing the trustworthiness of the system [9].

For AI models to be trusted in clinical settings, they must not only be accurate but also reliably indicate when their predictions may be unreliable. Such uncertainty can stem from the inherent noise of biological signals and patient variability, or from the model's own knowledge gaps due to limited training data [10, 11]. In practice, these challenges manifest as predictive uncertainty, which can be classified into three main types [5]: in-domain uncertainty for data drawn from the same distribution as the training set [12]; domain shift uncertainty, where the data is shifted from the training distribution [13, 14]; and out-of-distribution (OOD) uncertainty, where the data originate from an unknown distribution [15]. Effectively managing data shifts and OOD data is critical to ensure robust and safe deployment [16]. To achieve this, methods like Bayesian inference and ensembles are used to create well-calibrated models, where the output probability distribution aligns with the true likelihood of outcomes, making the model's confidence explicitly interpretable [17]. The challenges of uncertainty are especially pronounced in neurophysiology, where signals are notoriously complex and noisy.

Electroencephalography (EEG) is a prime example, which is a non-invasive method for measuring the brain's electrical activity from scalp electrodes. The recorded signals are multivariate time series that reflect the summation of postsynaptic potentials from neuronal populations, shaped by the orientation and proximity of the underlying signal sources [18]. However, it is highly susceptible to intrinsic and extrinsic artefacts, with a low signal-to-noise ratio [19]. DL and machine learning have been successfully applied to EEG data for tasks such as brain-computer interfaces, sleep staging, routine EEG interpretation [20] and prediction of cognitive states [19]. Un-

certainty estimation in the context of EEG and deep learning has been proposed for several purposes. It has been used to assist in artefact removal from EEG signals [21] and to improve performance in brain–computer interfaces [22]. In seizure detection, uncertainty estimation has been applied to increase confidence in patient-level predictions [23]. In the domain of sleep staging, expressing uncertainty as a singular metric has been shown to enhance model transparency [24, 25], with a significant correlation observed between uncertainty estimates and model performance [24]. [26] extends this further by addressing uncertainty in motor-imagery tasks and aim to reduce uncertainty with the adaptive data augmentation approach [26]. Furthermore, previous work also demonstrated the use of uncertainty estimation in sex classification from EEG, highlighting how ensemble-based methods can improve predictive performance [27]. These applications demonstrate that quantifying uncertainty is a critical step toward developing more robust and trustworthy AI systems for clinical neurology.

In this study, we investigate how different ensemble strategies impact both predictive performance and uncertainty estimation in the classification of Normal, Mild Cognitive Impairment (MCI), and Dementia from resting-state EEG recordings. While grounded in a clinically relevant classification task, this work focuses on methodological evaluation of model performance and uncertainty. Our focus is on assessing model behaviour under realistic clinical conditions, including heterogeneous populations, real-world EEG variability, and distributional shifts. Specifically, we evaluate ensemble methods, including independently trained deep ensembles, checkpoint-based approaches, and methods such as Monte Carlo dropout and Stochastic Weight Averaging Gaussian, under three conditions: (1) a held-out in-distribution test set, (2) three relevant OOD datasets, and (3) controlled, EEG-specific dataset shifts constructed by systematically altering signal characteristics.

Through this comprehensive evaluation, we aim to answer two key questions: (1) which ensemble methods yield the most reliable predictions and well-calibrated uncertainty estimates across different datasets, and (2) how uncertainty relates to predictive performance across increasing levels of dataset shift. By examining these questions in the context of clinically relevant EEG classification, this study contributes to the development of more robust and trustworthy AI systems for clinical decision support.

2. Methods and materials

2.1. Dataset

2.1.1. Training dataset The EEG dataset used is the South Korean CAUEEG dataset [28], which includes 1,176 subjects clinically classified into three categories: normal, MCI and dementia. The dataset is pre-divided into 80% training, 10% validation, and 10% test sets, a split adopted in this study for consistency and comparability. The cohort had an age range of 23-96 years (mean = 71.13 ± 9.81). The dataset consists of EEG recordings in resting state, with segments associated with photic stimulation. Only sections preceding the photic stimuli were analysed, starting from the first eyes-closed marker. The preprocessing pipeline included bandpass filtering between 1.0 Hz and 45.0 Hz, re-referencing to the average, and removing line noise using a notch filter. EEG data were segmented into non-overlapping 30-second epochs. Autoreject [29] was then applied to remove epochs with

excessive noise and artefacts and to repair channels. To improve model generalisation, data augmentation techniques were applied, limited to those that did not overlap with dataset shift, including time reversal, sign flipping, and smooth time masking (mask length = 20). Four epochs were selected per subject, ensuring maximal spread across the available epochs to increase variance. The final prediction for each subject was obtained by averaging the softmax outputs of the predictions across the four epochs.

In addition to EEG, the subject's age was also used as input to the model due to its impact on performance [28], since age is a significant predictor of dementia [30]. Age data were normalised using min-max scaling, and random Gaussian noise ($\mu = 0, \sigma = 0.2$) was added with a probability of 0.5 to account for variability and improve model generalisation.

2.1.2. OOD datasets Three OOD datasets were used: Militadous [31]‡, MPI [32], and TDBRAIN [33]. All OOD datasets were subjected to the same pre-processing pipeline as CAUEEG, and only the resting state EEG data were selected. First, a subset of channels was selected to match the 19 channels in the CAUEEG. In TDBRAIN and MPI, channels T7, T8, P7 and P8 were renamed to account for the absence of channels T3, T4, T5 and T6, as they cover the same regions, reflecting differences in naming conventions between the 10–20 and 10–10/10–5 system. Additionally, in MPI, some subjects were excluded due to missing electrodes. The labels varied between the datasets. The Miltiadous dataset comprises 88 participants (44 female, 44 male) aged 49–79 years (mean = 66.17 ± 7.32), clinically stratified into Alzheimer's disease ($n = 36$), frontotemporal dementia ($n = 24$), and healthy controls ($n = 30$). For the current analysis, the two dementia groups were merged into a single 'dementia' class. The MPI dataset includes 131 healthy subjects (90 male, 41 female) aged 22.5–77.5 years (mean = 38.26 ± 20.02), after excluding participants with missing channels. The TDBRAIN dataset provided 27 healthy control participants (15 male, 12 female), aged 18.3–55.3 years (mean = 29.02 ± 11.26), selected from a larger clinical cohort referred for neurophysiological evaluation. Only individuals with complete data were kept. In addition to EEG, age was also used as an input feature for the OOD datasets.

‡ For practical reasons, we use the name of the first author to refer to this dataset throughout the paper

2.2. Designing dataset shifts

Multiple synthetic shifts were developed to simulate targeted perturbations in EEG data, enabling the evaluation of model performance and uncertainty under controlled distributional shifts. The objective was to introduce controlled, gradual shifts that preserve the complex signal integrity of EEG, in order to systematically evaluate the model's response to specific types of signal alteration. This approach facilitates the evaluation of model robustness under varying degrees of dataset shift.

This framework was defined by a shift intensity parameter x. Notably, the interpretation of x depends on the type of dataset shift being applied. For static shifts, such as channel interpolation and channel rotation, the shift intensity x determined the proportion of EEG channels that were affected, with values defined as $x \in \{0.1, 0.25, 0.5, 0.75, 0.9, 1.0\}$. In contrast, for magnitude-based shifts, such as additive Gaussian noise, the parameter x specified the severity of the perturbation, for exam-

ple by controlling the standard deviation of the noise applied.

To formalise the notation, let $S_i(t)$ represent the signal of channel i at time t. Importantly, for the magnitude-based shifts, the noise applied to each channel is sampled independently, ensuring that the variations are channel-specific rather than global. The noise magnitude for dataset shifts is drawn from the distribution:

$$\mathcal{R}(x) := U[0.75x, 1.25x] \tag{1}$$

where $U[a, b]$ is the uniform distribution in the interval [a,b] and x specifies the intended noise magnitude. This introduces moderate variability around x, producing more heterogeneous but bounded perturbations. In some shift types, the direction of the noise was also randomised, represented by:

$$\pm\mathcal{R}(x) \tag{2}$$

This represents a 50% probability of positive or negative perturbation per instance. Determining an appropriate value for x is non-trivial. For each type of shift, several values of x were empirically tested to explore their effect on the signal. Some of the implemented shift types have previously been proposed as data augmentation strategies [34].

2.2.1. Gaussian noise In EEG recordings, noise can originate from multiple sources, including sensor defects, environmental interference, or physiological artefacts (e.g. muscle activity). Let $\epsilon_i(t)$ denote Gaussian noise sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$. The augmented signal $\widetilde{S}_i(t)$ is defined by

$$\widetilde{S}_i(t) = S_i(t) + \epsilon_i(t), \tag{3}$$

2.2.2. Interpolation To simulate sensor dropout or channel malfunction, a subset of EEG channels was temporarily removed and then reconstructed using interpolation based on a Minimum Norm Estimate [35]. At maximum shift intensity, all channels except two (Cz and Pz) are removed and interpolated.

2.2.3. Bandstop filtering The signals were bandstop filtered to remove specific frequency bands. The following frequency bands were investigated: delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz) and beta (12–30 Hz). Finally, a low-pass filter, filtering out gamma frequencies (>30Hz). This shift is the only shift which does not have a shift intensity.

2.2.4. Channel shuffling To simulate variations in electrode placement, a subset of EEG channels was randomly shuffled. The selected channels are randomly reordered, with no channel remaining in its original position, and the same permutation is applied to all subjects. This shift introduces controlled spatial variability while preserving the temporal structure of the signals.

2.2.5. Amplitude modulation To simulate the variability in signal gain and electrode sensitivity, the amplitude of each EEG channel was modulated by a scalar factor. Let $k_i \sim \mathcal{R}(k)$ be a modulation factor of channel i.

$$\widetilde{S}_i(t) = k_i \cdot S_i(t). \tag{4}$$

2.2.6. Phase shift To introduce variability into the dynamics of EEG signals, we applied a phase shift to each channel. Let $\phi_i \sim \pm\mathcal{R}(\phi)$ be the phase change in radians, and let $\mathcal{H}\{S_i(t)\}$ denote the Hilbert transform of $S_i(t)$.

$$\widetilde{S}_i(t) = \Re\left\{|\mathcal{H}\{S_i(t)\}| \cdot e^{j(\theta_i(t)+\phi_i)}\right\} \quad (5)$$

2.2.7. Circular shift To introduce variability in the temporal alignment of the EEG signals, we applied a circular change to each channel, where $k_i \sim \pm\mathcal{R}(k)$, and T is the total number of time samples.

$$\widetilde{S}_i(t) = S_i\Big((t + k_i) \bmod T\Big). \quad (6)$$

2.2.8. Slow drift In order to simulate slow-varying offset fluctuations (e.g., due to perspiration or electrode impedance changes), a slow drift was added to each EEG channel. The drift function $\Delta_i(t)$ is constructed as:

$$\Delta_{b_i}(t) = A_i \sum_{k=1}^{K_i} \alpha_{ik} \sin(2\pi f_{ik}\, t + \phi_{ik}), \quad (7)$$

where $A_i \sim \mathcal{R}(\mathcal{A})$ is the maximum amplitude and a $K_i \sim \mathcal{R}(\mathcal{K})$ is the number of sinusoidal waves, $\alpha_{ik} \sim U[0.5, 1.0]$ is a weighting factor for the k-th sinusoid, $f_{ik} \sim U[0.01, 0.2]$ is the frequency and $\phi_{ik} \sim U[0, 2\pi]$ is the phase of each sinusoid. The drifted signal is then defined as:

$$\widetilde{S}_i(t) = S_i(t) + \Delta_{b_i}(t). \quad (8)$$

2.2.9. Non-linear time warping To introduce temporal distortions in EEG signals, a non-linear time-warping transformation was applied. We define the warp function $w(t)$ as:

$$w_i(t) = t + \Delta_w(t), \quad (9)$$

where $\Delta_{w_i}(t)$ is a sinusoidal displacement:

$$\Delta_{w_i}(t) = \alpha_i \sin\Big(2\pi f_i \tfrac{t}{T} + \phi_i\Big) \times \big(T\,\gamma_i\big). \quad (10)$$

Here, $\alpha_i \sim \mathcal{R}(\alpha)$, T is the total number of time points $f_i \sim U[0.01, 0.2]$ is the random frequency, $\phi_i \sim U[0, 2\pi]$ is a random phase change and $\lambda_i \sim [0.5, 1.0]$ is a randomly sampled scaling factor. Since $w_i(t)$ often yields non-integer indices, the final warped signal $\widetilde{S}_i(t)$ is obtained by linearly interpolating $S_i(t)$ at these warped indices:

$$\widetilde{S}_i(t) = \mathrm{interp}\big(S_i,\ w_i(t)\big), \quad (11)$$

2.2.10. Peak shift To simulate a shift in the alpha band (8-12 Hz), we first isolate the alpha component $\alpha_i(t)$ from an EEG signal $S_i(t)$ using band-pass filtering. We define the residual outside the alpha band as:

$$\beta_i(t) = S_i(t) - \alpha_i(t). \quad (12)$$

Next, we compute the analytic signal of $\alpha_i(t)$ using the Hilbert transform:

$$\alpha_{\mathrm{ia}}(t) = \mathcal{H}\{\alpha_i(t)\}. \quad (13)$$

Then a frequency shift $\Delta_{f_i} \sim \pm\mathcal{R}(\Delta_{f_i})$ was applied to the analytic signal by multiplying it with a complex exponential and taking the real part:

$$\widetilde{\alpha}_i(t) = \mathrm{Re}\left[\alpha_{\mathrm{ia}}(t) e^{j2\pi\Delta_{f_i} t}\right]. \quad (14)$$

Finally, we reconstruct the shifted EEG signal by adding $\widetilde{\alpha}_i(t)$ back to the residual $\beta_i(t)$:

$$\widetilde{S}_i(t) = \beta_i(t) + \widetilde{\alpha}_i(t). \quad (15)$$

2.3. Model

The primary DL model used in this study is the InceptionNetwork [36], which has shown strong performance in EEG classification tasks [37, 38, 27, 39]. All ensembles were constructed using this single architecture, and for ensemble types that require repeated training runs, the

ensemble size was fixed at five, consistent with previous studies [40]. For ensemble methods requiring only a single training run, the ensemble size was set to 20 to balance predictive performance and computational cost. Hyperparameter tuning was performed using the Optuna framework [41], employing the Tree-structured Parzen Estimator (TPE) algorithm [42]. A total of 125 trials were conducted per ensemble type to ensure fairness between the ensemble methods. The randomly initialised weight ensemble was excluded from tuning due to its lack of adjustable parameters.

Base model The base model consists of a six-layer CNN with 20 units per layer and a maximum kernel size of 60. It includes two fully connected layers with 256 and 128 neurons, employing batch normalisation and dropout (rate: 0.5) to enhance generalisation. The ensemble specific parameters and changes in the model can be found in the github repository §.

Temperature scaling This study employed temperature scaling, a post-hoc calibration technique proposed by Guo et al. [43], which rescales the logits by a scalar temperature parameter to adjust the confidence of the softmax outputs without altering the predicted class labels. Temperature scaling was applied to the aggregated logits of the ensemble, and the temperature values were optimised using the validation set.

2.3.1. Deep Ensemble The Deep Ensemble [40] is trained by initialising multiple models independently using different random seeds, allowing each model to explore distinct regions of the parameter space.

2.3.2. Bagging Ensemble The Bagging Ensemble [44] approach works by randomly drawing a subset of the training set with replacement and using this subset to train a classifier. By repeating this process multiple times, an ensemble of classifiers trained on different subsamples of the data is created. This ensemble used smaller versions of the base models as it is trained using less data.

2.3.3. Depth Ensemble The Depth Ensemble takes advantage of architectural variation by altering the number of inception modules within the InceptionNetwork. By varying the number of Inception modules, the Depth Ensemble creates models with different capacities to extract and represent features.

2.3.4. Augmentation Ensemble The augmentation ensemble technique differs from the deep ensemble by also using different data augmentation techniques per model in the ensemble.

2.3.5. Monte Carlo Dropout The Monte Carlo dropout (MCD) [45] approximates a Bayesian method by randomly dropping neurons during inference time. This *sampling* with multiple forward passes of different architectures with different neurons turned off generates a computationally efficient ensemble. In this study, 50 stochastic forward passes were used during inference.

2.3.6. Snapshot Ensemble The snapshot ensemble [46] generates M models in a single training session. This method rapidly changes the learning rate using a cyclic cosine

§ `https://github.com/matstveter/eeg_dl_uncertainty`

annealing schedule [47]. This approach enables the model to explore multiple local minima, with the schedule periodically increasing the learning rate to escape the current minima.

2.3.7. Stochastic Weight Averaging Gaussian Stochastic Weight Averaging Gaussian (SWAG) [48] extends Stochastic Weight Averaging (SWA) [49] by modeling the distribution of weights in parameter space. Instead of averaging weights at selected checkpoints, SWAG estimates a Gaussian distribution over the weights by collecting the mean and low-rank covariance from multiple points during training. In our implementation, weights were collected from a partially trained model to estimate the SWAG distribution. At test time, 50 weight samples were drawn to form an ensemble, and their predictions were averaged.

2.3.8. Fast Geometric Ensemble The Fast Geometric Ensemble (FGE) [50] technique is similar to the snapshot ensemble [46] in the way models are saved during one iteration of the training and similar to SWA by starting with a pre-trained model. It differs from SE by using a piecewise linear cyclical learning rate schedule rather than the cosine schedule, and, in addition, the cycles are usually smaller in magnitude.

2.3.9. Ensemble taxonomy The ensemble methods can be grouped into three categories based on how they create and maintain diversity. Multi-run ensembles, including Deep Ensemble, Bagging, Augmentation, and Depth, train multiple independently initialised models. Checkpoint-based methods, such as Snapshot and FGE, extract ensemble members from intermediate stages of a single training run. Sampling-based methods, including MCDropout and SWAG, derive diversity by sampling during inference using a single trained model in the case of MCD and drawing weights from a learnt posterior approximation in the case of SWAG.

2.4. Performance metrics

For this study, Area-Under-the-curve (AUC) [51] is used to compare the performance of the models. The AUC has a value between 0.0 and 1.0, with 0.5 representing random guessing. For multiple classes, the One-vs-Rest (Ovr) strategy is used, where the AUC is computed for each class against all others and then averaged using a weighted approach to provide an overall performance measure. For per-class AUC, Ovr is applied without weighting, where AUC is calculated separately for each class. For comparisons involving OOD datasets with fewer classes, recall and precision are reported per class. Recall reflects the proportion of correctly identified positives, while precision indicates the accuracy of positive predictions.

2.5. Uncertainty metrics

Brier score The Brier score [52] is a strictly proper scoring rule that evaluates the quality of probabilistic predictions by jointly capturing both accuracy and calibration. It measures how close the predicted probability distribution is to the actual outcome. A lower score reflects better performance, indicating that the model not only predicts the correct class more often but also assigns appropriate levels of confidence to its predictions. In other words, the Brier score penalizes both incorrect predictions and over- or underconfident probability estimates. A perfectly calibrated and accurate model would achieve a score of 0. The Brier score for multiclass classification is defined as:

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(p_{i,k} - o_{i,k})^2$$

where N is the number of instances, K is the number of classes, $p_{i,k}$ is the predicted probability for instance i belonging to class k and $o_{i,k}$ is the actual outcome which is 1 if the instance i belongs to class k, and 0 otherwise.

Expected Calibration Error (ECE) The ECE quantifies how well the predicted probabilities match the observed frequencies, providing information on model calibration. Predictions are grouped into bins B according to their confidence scores. For each bin b, the accuracy and confidence are calculated as:

$$\text{Accuracy}(b) = \frac{\sum_{i \in b} \mathbb{1}(\hat{y}_i = y_i)}{|b|} \tag{16}$$

$$\text{Confidence}(b) = \frac{\sum_{i \in b} \max_k p_{i,k}}{|b|}, \tag{17}$$

where $\max_k p_{i,k}$ is the highest predicted probability for sample i, and $\mathbb{1}(\hat{y}_i = y_i)$ is an indicator function that equals 1 if the predicted class $\hat{y}_i = \arg\max_k p_{i,k}$ matches the true labely_i, and 0 otherwise. The ECE is then calculated as the weighted average of the absolute difference between accuracy and confidence across bins:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|b|}{N} \left|\text{Accuracy}(b) - \text{Confidence}(b)\right|, \tag{18}$$

where $|b|$ is the number of predictions in bin b, and N is the total number of samples. A lower ECE indicates better calibration, meaning that the predictions align more closely to the observed outcomes. The number of bins was set to 10.

3. Results

3.1. In-distribution test performance

Table 1 reports several metrics evaluating the performance of the different ensemble methods on the test set. Starting with overall AUC, we observe that the multi-run ensembles, including MCD, demonstrate strong predictive performance. In contrast, checkpoint ensembles perform significantly worse, with non-overlapping confidence intervals. SWAG falls in between these groups. For accuracy, a metric more sensitive to class imbalance, all multi-run models achieve over 60%, while checkpoint and sampling models fall below this threshold. Notably, SWAG displays high run-to-run variability. Looking at the per-class AUC, all models perform best in classifying normal and dementia cases, with the top AUC values exceeding 0.90. However, MCI classification remains challenging, with the best AUC reaching only 0.71.

For the Normal class, the recall of multi-run ensembles is high ($\geq$ 0.90), moderately high for Dementia ($\geq$ 0.68), and low for MCI (0.27–0.41). Precision follows the exact ordering, with about 0.70 for Normal and Dementia and around 0.60 for MCI. Checkpoint ensembles show markedly lower recall for Normal, reduced recall for Dementia, and similar recall for MCI; their precision is comparable to the multi-run group for Normal but lower for both Dementia and MCI. MCD confidence intervals overlap those of the multi-run models in every class, yet its average recall for Normal and Dementia is slightly lower, and the intervals themselves are wider. For precision, MCD has a similar performance to that of multi-run ensembles. SWAG displays the greatest instability, its recall, and precision intervals are extremely broad, with

Ensemble Type	AUC	Accuracy (%)	$\text{Recall}_{\text{class}}$	$\text{Precision}_{\text{class}}$	$\text{AUC}_{\text{class}}$
Deep Ensemble	0.84 ± 0.01	67.18 ± 2.07	N: 0.90 ± 0.06 M: 0.41 ± 0.10 D: 0.68 ± 0.07	N: 0.71 ± 0.07 M: 0.59 ± 0.02 D: 0.69 ± 0.06	N: 0.91 ± 0.01 M: 0.70 ± 0.02 D: 0.91 ± 0.01
Bagging	0.83 ± 0.02	65.13 ± 3.93	N: 0.92 ± 0.03 M: 0.27 ± 0.12 D: 0.77 ± 0.12	N: 0.68 ± 0.06 M: 0.56 ± 0.12 D: 0.66 ± 0.05	N: 0.90 ± 0.01 M: 0.68 ± 0.04 D: 0.91 ± 0.01
Depth	0.83 ± 0.02	68.21 ± 2.75	N: 0.92 ± 0.06 M: 0.40 ± 0.12 D: 0.71 ± 0.08	N: 0.70 ± 0.04 M: 0.63 ± 0.06 D: 0.69 ± 0.07	N: 0.91 ± 0.02 M: 0.69 ± 0.03 D: 0.90 ± 0.01
Augmentation	0.84 ± 0.02	67.69 ± 2.30	N: 0.92 ± 0.04 M: 0.33 ± 0.08 D: 0.78 ± 0.06	N: 0.68 ± 0.04 M: 0.62 ± 0.07 D: 0.71 ± 0.04	N: 0.91 ± 0.01 M: 0.71 ± 0.04 D: 0.92 ± 0.02
MCD	0.80 ± 0.03	59.66 ± 5.69	N: 0.86 ± 0.12 M: 0.29 ± 0.15 D: 0.61 ± 0.37	N: 0.65 ± 0.10 M: 0.49 ± 0.08 D: 0.66 ± 0.13	N: 0.87 ± 0.02 M: 0.66 ± 0.02 D: 0.89 ± 0.04
Snapshot	0.71 ± 0.03	52.82 ± 5.59	N: 0.64 ± 0.11 M: 0.39 ± 0.06 D: 0.56 ± 0.07	N: 0.65 ± 0.02 M: 0.40 ± 0.08 D: 0.53 ± 0.05	N: 0.80 ± 0.02 M: 0.53 ± 0.04 D: 0.80 ± 0.02
SWAG	0.77 ± 0.04	40.00 ± 16.46	N: 0.38 ± 0.52 M: 0.16 ± 0.43 D: 0.74 ± 0.54	N: 0.76 ± 0.22 M: 0.28 ± 0.54 D: 0.30 ± 0.30	N: 0.86 ± 0.04 M: 0.58 ± 0.05 D: 0.87 ± 0.03
FGE	0.74 ± 0.01	57.91 ± 4.48	N: 0.76 ± 0.03 M: 0.37 ± 0.07 D: 0.58 ± 0.07	N: 0.67 ± 0.04 M: 0.43 ± 0.07 D: 0.58 ± 0.04	N: 0.84 ± 0.01 M: 0.56 ± 0.02 D: 0.83 ± 0.01

Table 1: Performance on the holdout test set with 117 subjects, in predicting between the classes: Normal, MCI and Dementia. The performance is calculated across 5 runs per ensemble with mean and 95% confidence interval. Where N=Normal, M=MCI and D=Dementia for the 3 right columns.

dementia recall dipping to near zero in some runs and reaching 1.0 in others, underscoring substantial run-to-run variability.

3.2. Out-of-Distribution

3.2.1. Performance Table 2 presents class-wise recall and precision for three OOD cohorts. In Militadous, the multi-run ensembles maintain strong recall for the normal class (0.82-0.89) and moderate recall for dementia (0.52-0.57). Their precision remains balanced, about 0.50 for Normal and above 0.93 for Dementia. Depth is slightly below the average of the other multi-run models, but its confidence intervals overlap. MCD shows similar behaviour for Normal but lower and more variable recall for Dementia, with precision similar to the multi-run ensembles. The checkpoint ensembles lose much more recall and precision, and

Ensemble	**Militadous**		**MPI**	**TDBRAIN**
	$\text{Recall}_{\text{class}}$	$\text{Precision}_{\text{class}}$	$\text{Recall}_{\text{class}}$	$\text{Recall}_{\text{class}}$
Deep Ensemble	N: 0.86 ± 0.12	N: 0.49 ± 0.02	N: 0.96 ± 0.04	N: 1.00 ± 0.00
	D: 0.52 ± 0.07	D: 1.00 ± 0.00		
Bagging	N: 0.82 ± 0.13	N: 0.50 ± 0.03	N: 0.96 ± 0.03	N: 0.99 ± 0.02
	D: 0.53 ± 0.08	D: 0.96 ± 0.02		
Depth	N: 0.73 ± 0.16	N: 0.52 ± 0.05	N: 0.95 ± 0.04	N: 0.98 ± 0.03
	D: 0.54 ± 0.09	D: 0.93 ± 0.04		
Augmentation	N: 0.89 ± 0.14	N: 0.52 ± 0.03	N: 0.98 ± 0.02	N: 1.00 ± 0.00
	D: 0.57 ± 0.07	D: 0.98 ± 0.03		
MCD	N: 0.81 ± 0.10	N: 0.49 ± 0.10	N: 0.91 ± 0.05	N: 0.93 ± 0.08
	D: 0.46 ± 0.23	D: 0.96 ± 0.03		
Snapshot	N: 0.47 ± 0.10	N: 0.42 ± 0.05	N: 0.37 ± 0.05	N: 0.36 ± 0.06
	D: 0.34 ± 0.05	D: 0.74 ± 0.05		
SWAG	N: 0.50 ± 0.46	N: 0.65 ± 0.26	N: 0.77 ± 0.28	N: 0.74 ± 0.38
	D: 0.66 ± 0.56	D: 0.64 ± 0.47		
FGE	N: 0.51 ± 0.07	N: 0.44 ± 0.02	N: 0.38 ± 0.03	N: 0.31 ± 0.04
	D: 0.35 ± 0.04	D: 0.74 ± 0.04		

Table 2: Comparison of precision and recall calculated per class in the 3 OOD datasets. The performance is calculated across 5 runs per ensemble with mean and 95% confidence interval. Where N=Normal and D=Dementia.

SWAG shows the widest run-to-run variability, making its performance unreliable.

The single-class cohorts MPI and TDBRAIN follow the same ranking pattern. All multi-run methods recall almost every normal subject (about 0.95-1.00). MCD is slightly lower on average, but its intervals overlap. Checkpoint ensembles miss a substantial share of normals, and while SWAG average recall is higher, its confidence intervals are so wide that its reliability remains doubtful.

3.2.2. Uncertainty calibration Table 3 summarises the Brier score and ECE for each ensemble on the test set and three OOD cohorts. On the test set, the multi-run ensembles record the lowest Brier values (0.45–0.47) and mid-range ECE error (0.10–0.13). The MCD Brier score is slightly higher and its ECE is slightly lower, but both confidence intervals overlap those of the multi-run group. The checkpoint ensembles have noticeably higher Brier scores than the multi-run methods, while their ECE is on average lower, though again with overlapping intervals. SWAG stands out with the highest Brier and ECE values and the widest confidence intervals. In the Militadous cohort, every ensemble shows a higher Brier score than on the test set, except SWAG, but with much wider confidence intervals, and the Augmentation ensemble, which remains stable (0.45 to 0.46). ECE error is largely stable for the multi-run methods with ECE near 0.10. In contrast, the checkpoint and sampling methods all see noticeably higher ECE values and broader intervals than on the test set. In

(a) Test (Normal/MCI/Dementia)

Ensemble	Brier	ECE
Deep Ensemble	0.45 ± 0.01	0.10 ± 0.02
Bagging	0.47 ± 0.01	0.11 ± 0.03
Depth	0.47 ± 0.00	0.13 ± 0.06
Augmentation	0.45 ± 0.01	0.10 ± 0.04
MCD	0.51 ± 0.05	0.08 ± 0.02
Snapshot	0.57 ± 0.03	0.08 ± 0.05
SWAG	0.64 ± 0.08	0.17 ± 0.08
FGE	0.55 ± 0.01	0.08 ± 0.04

(b) Militadous (Normal/Dementia)

Ensemble	Brier	ECE
Deep Ensemble	0.52 ± 0.02	0.10 ± 0.03
Bagging	0.50 ± 0.06	0.08 ± 0.02
Depth	0.54 ± 0.05	0.11 ± 0.05
Augmentation	0.46 ± 0.03	0.10 ± 0.03
MCD	0.58 ± 0.15	0.14 ± 0.07
Snapshot	0.75 ± 0.04	0.18 ± 0.08
SWAG	0.53 ± 0.13	0.24 ± 0.07
FGE	0.75 ± 0.01	0.13 ± 0.03

(c) MPI (Normal)

Ensemble	Brier	ECE
Deep Ensemble	0.11 ± 0.05	0.17 ± 0.02
Bagging	0.13 ± 0.06	0.20 ± 0.04
Depth	0.18 ± 0.08	0.24 ± 0.05
Augmentation	0.11 ± 0.05	0.18 ± 0.05
MCD	0.24 ± 0.09	0.25 ± 0.03
Snapshot	0.71 ± 0.02	0.18 ± 0.05
SWAG	0.40 ± 0.24	0.24 ± 0.15
FGE	0.73 ± 0.01	0.16 ± 0.02

(d) TDBRAIN (Normal)

Ensemble	Brier	ECE
Deep Ensemble	0.06 ± 0.03	0.16 ± 0.04
Bagging	0.08 ± 0.02	0.19 ± 0.03
Depth	0.14 ± 0.06	0.25 ± 0.06
Augmentation	0.06 ± 0.03	0.16 ± 0.06
MCD	0.26 ± 0.10	0.32 ± 0.02
Snapshot	0.70 ± 0.02	0.24 ± 0.02
SWAG	0.42 ± 0.25	0.28 ± 0.15
FGE	0.72 ± 0.02	0.28 ± 0.06

Table 3: Uncertainty comparison of ensemble methods (mean ±95% CI over 5 runs) calculated across the test set and three OOD datasets.

the single-class MPI and TDBRAIN cohorts, Brier scores decrease for the multi-run ensembles and the sampling methods relative to the test set, whereas the checkpoint ensembles exhibit higher Brier scores. ECE values increase for every ensemble in both cohorts relative to their test set values.

3.3. Dataset shifts

3.3.1. Performance and uncertainty calibration We evaluated the models performance using AUC, and for uncertainty analysis, we investigated the Brier score and ECE.

Amplitude shifts (sub-figures 1a, 1b, and 1c) reveal three distinct patterns. Checkpoint ensembles maintain almost flat curves: their AUC, Brier score, and ECE remain stable across the full range of shift intensities. The sampling methods show worse performance, and MCD shows higher Brier and ECE at the lowest shift intensities. Multi-run ensembles show the highest performance around ($\times 0.001$–$\times 2$), but their performance drops below the checkpoint models once the signal is amplified beyond about $\times 2$. Between $0.25\times$ and $2\times$ scaling, the multi-run ensembles show lower Brier scores than the checkpoint models, but outside this

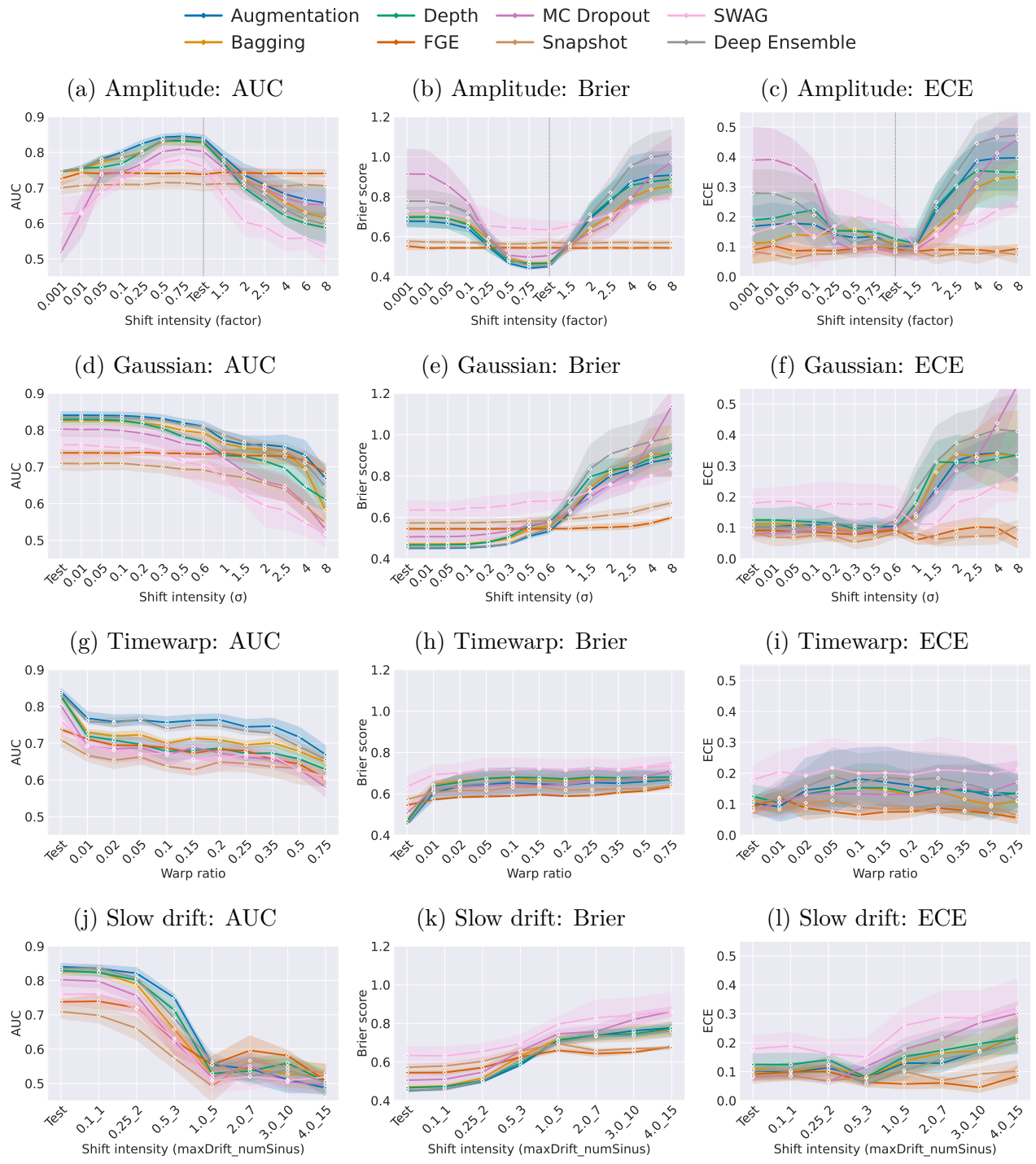

Figure 1: Illustrating ensemble performance, uncertainty and calibration under four dataset shifts. Columns plot three metrics, AUC, Brier score and ECE, against shift intensity. The rows (top to bottom) represent shift types: (1) amplitude scaling (signal multiplied by intensity), (2) Gaussian noise (added noise), (3) time warp (local compression/expansion), and (4) slow drift (added low-frequency sinusoidal waves). Ensemble methods are labelled in the top of the figure.

window, their Brier scores are generally higher. Across the full amplitude range, they also exhibit higher ECE values and broader confidence intervals. Across all methods, the Augmentation ensemble achieves marginally better performance, although the differences are not statistically significant.

Under Gaussian noise (sub-figures 1d, 1e, and 1f) all ensembles maintain baseline AUC until the noise standard deviation reaches about $\sigma = 0.2$. Beyond this point, the multi-run methods continue to rank highest, with Augmentation and Deep ensembles remaining the best performers all the way to $\sigma = 8$ despite a gradual decline. They also maintain the lowest Brier scores up to roughly $\sigma = 0.6$. Checkpoint models are stable with FGE keeping both AUC and Brier score essentially flat until around $\sigma = 4$, whereas Snapshot performance declines at a rate similar to the multi-run ensembles, but with a stable Brier score. The sampling methods are initially lower in performance than multi-run but higher than checkpoint ensembles, following the multi-run curves in both AUC and Brier. The ECE is stable for all ensembles until about $\sigma = 0.6$ beyond that, ECE rises sharply for the multi-run and sampling methods while remaining nearly constant for the checkpoint models.

Timewarp shifts (sub-figures 1g, 1h, and 1i) result in an immediate and substantial drop in performance for all ensembles, even at the lowest shift intensity. Performance continues to decline as the shift increases for most ensembles, except Augmentation and Deep ensembles, which keep the performance at around 0.2. The Brier scores reflect this trend, with the largest increase occurring after the first shift, followed by a more gradual rise. The checkpoint ensembles, while starting with a slightly higher baseline Brier score, show a flatter trajectory across shift levels. ECE values remain relatively stable on average for all models, but with wider confidence intervals.

Slow drift (sub-figures 1j, 1k, and 1l) causes a gradual decline in performance below 0.25 Hz / 2 sinusoidal components, followed by a sharp drop between 0.25 Hz and 1.0 Hz / 5 sinus waves. While a larger grid of combinations was tested, only a representative subset is shown to highlight the overall trend. The augmentation ensemble seems to have a slower descent, at least until the mark of 1.0 Hz / 5 sinus waves. Beyond this point, model performance remains generally poor across all ensembles, with an AUC of around 0.5. The Brier scores follow a similar pattern, although the checkpoint models display a flatter curve after the 1.0 Hz / 5-wave mark. ECE values remain stable for most models up to 0.5 Hz / 3 sinus waves, after which calibration quality begins to degrade, again with checkpoint models maintaining lower ECE compared to the rest.

Phase shifts (sub-figures 2a, 2b and 2c) result in decreased performance across all shift levels, with the most pronounced drop occurring at phase offsets of $\pi/2$ and $3\pi/2$ (i.e., 90° and 270°). The relative ranking of ensemble performance seen on the test set remains largely consistent across all phase shifts, with the augmentation and deep ensemble performing slightly better throughout the shift intensities. The Brier score follows a similar pattern, reflecting increased uncertainty at $\pi/2$ and $3\pi/2$. In contrast, ECE remains relatively stable across most shifts and for all ensembles, with notable increases only at $\pi/2$ and $3\pi/2$.

Circular shifts (sub-figures 2d, 2e, and 2f) result in an immediate decline in performance across ensembles, with minor sensitivity to the magnitude of the shift. However, except for the

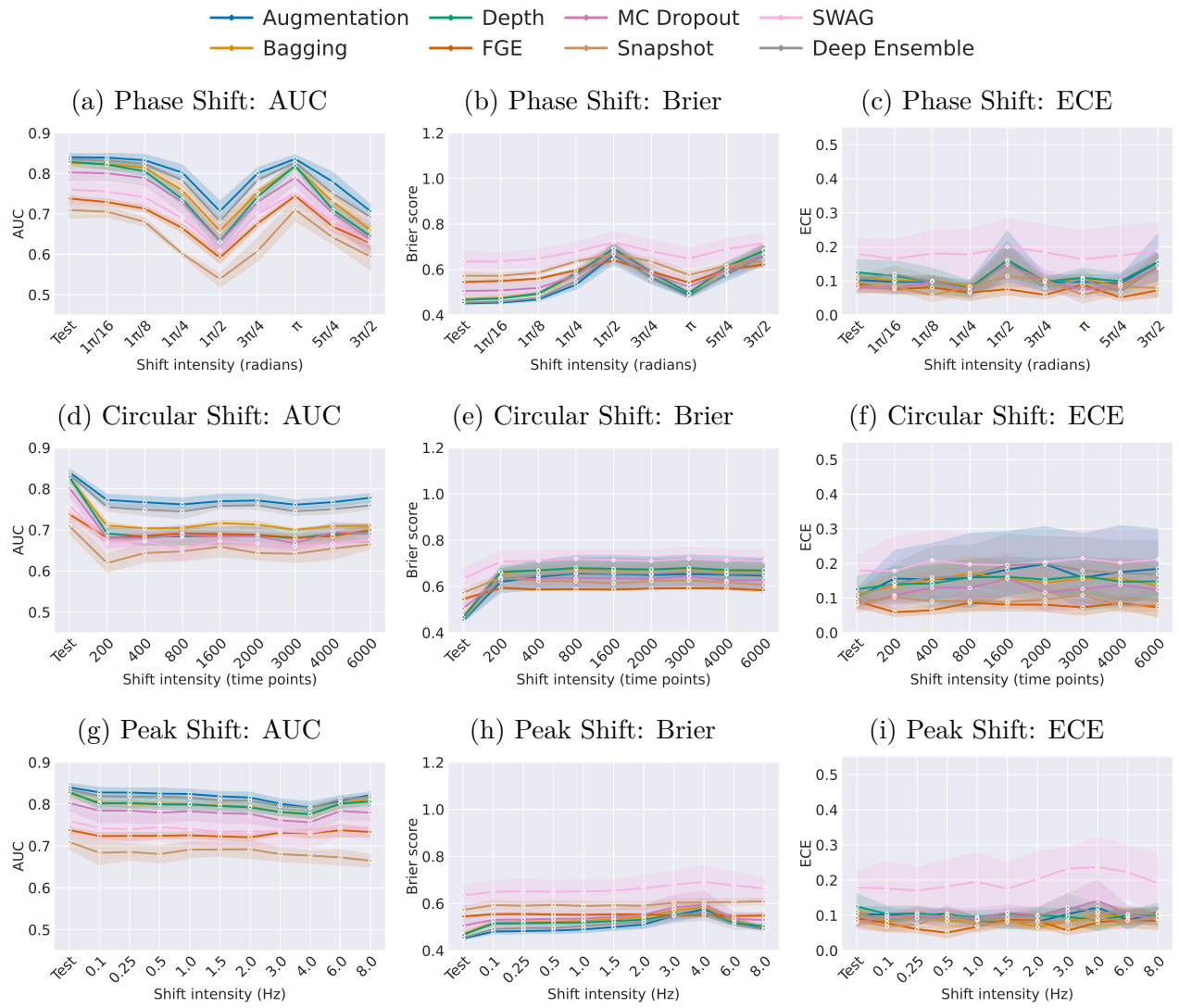

Figure 2: Illustrating ensemble performance, uncertainty and calibration under four dataset shifts. Columns plot three metrics, AUC, Brier score and ECE, against shift intensity. The rows (top to bottom) represent shift types: (1) phase shift (altering the signal's phase), (2) circular shift (rotating the signal in time), and (3) peak shift (displacing the alpha-band peak). Ensemble methods are labelled across the top of each column.

baseline, the deep ensemble and augmentation ensemble perform significantly better than the rest. The Brier score shows an increase in uncertainty from the first shift level, followed by a relatively stable plateau. For ECE, the average calibration remains largely stable across all ensembles, although a slow increase is observed for some. Notably, the confidence intervals for ECE are wide.

Peak shifts (sub-figures 2g, 2h, and 2i) result in a gradual, slow decline in performance, with more noticeable drop occurring around the 2–4 Hz range and the relative ordering of the ensemble performance is kept throughout the shift. A similar pattern is observed in the Brier score, showing a slight increase between

2 and 6 Hz, except for the checkpoint ensembles. ECE remains relatively stable, on average, throughout the entire frequency range, with a slight increase at 4.0 Hz for some of the ensembles.

Interpolation (sub-figures 3a, 3b and 3c) show that the performance remains stable until approximately 75% of channels are removed, after which it begins to decline moderately, with a sharp drop beyond 90%. The relative ordering of the performance of the ensemble remains stable. The Brier score remains stable until approximately 90% of the channels are removed, at which point it increases sharply. A similar trend is observed for ECE.

Channel rotation (sub-figures 3d, 3e, and 3f) leads to a steady decline in performance across all ensembles. The Deep and Augmentation ensembles start higher and decline more slowly than the others until rotation reaches roughly 75–90 percent. Their Brier curves rise at a correspondingly slower rate. Most ensembles exhibit a sharp increase in Brier score and ECE once rotation exceeds about 75 percent, whereas the checkpoint models retain slightly better calibration in that range. ECE remains nearly unchanged for all methods up to 50 percent rotation, and beyond this point the checkpoint ensembles continue to hold the most stable ECE values.

Bandstop shifts (sub-figures 4a, 4b, and 4c) show that SWAG maintains stable, with larger confidence intervals, across all frequency bands. In contrast, the other ensembles exhibit a significant decline in the theta band, and all except MC Dropout and Snapshot also show a drop in performance when the delta band is removed. Additionally, Bagging, Depth, MC Dropout, Snapshot, and Deep Ensemble show reduced performance in the beta band. These patterns are mirrored in the Brier scores, with clear increases in the delta, theta, and beta ranges. ECE remains relatively stable across all frequency bands and ensemble types.

3.4. Per-Class

Figure 5 shows the AUC for the normal class (Figure 5a) and the dementia class (Figure 5b) when specific frequency bands are removed. For the Normal class, removing either the theta or beta bands results in a marked drop in AUC, indicating that these bands contribute to the normal class classification. The dementia class displays a partially overlapping but distinct pattern: theta-band removal again leads to the largest performance drop across methods and larger than for the normal class, while delta and beta removal have a more modest impact, particularly in the augmentation-based ensemble.

4. Discussion

This study examined how different DL ensemble techniques and MCD classify Normal, MCI and Dementia from EEG and, more importantly, how well their uncertainty and calibration behave under distributional shift. Uncertainty and calibration were evaluated on the in-distribution test set, three OOD datasets and through a series of synthetic dataset shifts designed to represent feature-specific EEG noise. Across these conditions, multi-run ensembles consistently outperformed the other methods in terms of predictive accuracy while also maintaining a meaningful relationship between performance and uncertainty. These findings suggest that multi-run ensembles are better suited for robust and uncertainty-aware EEG classification in the face of data variability.

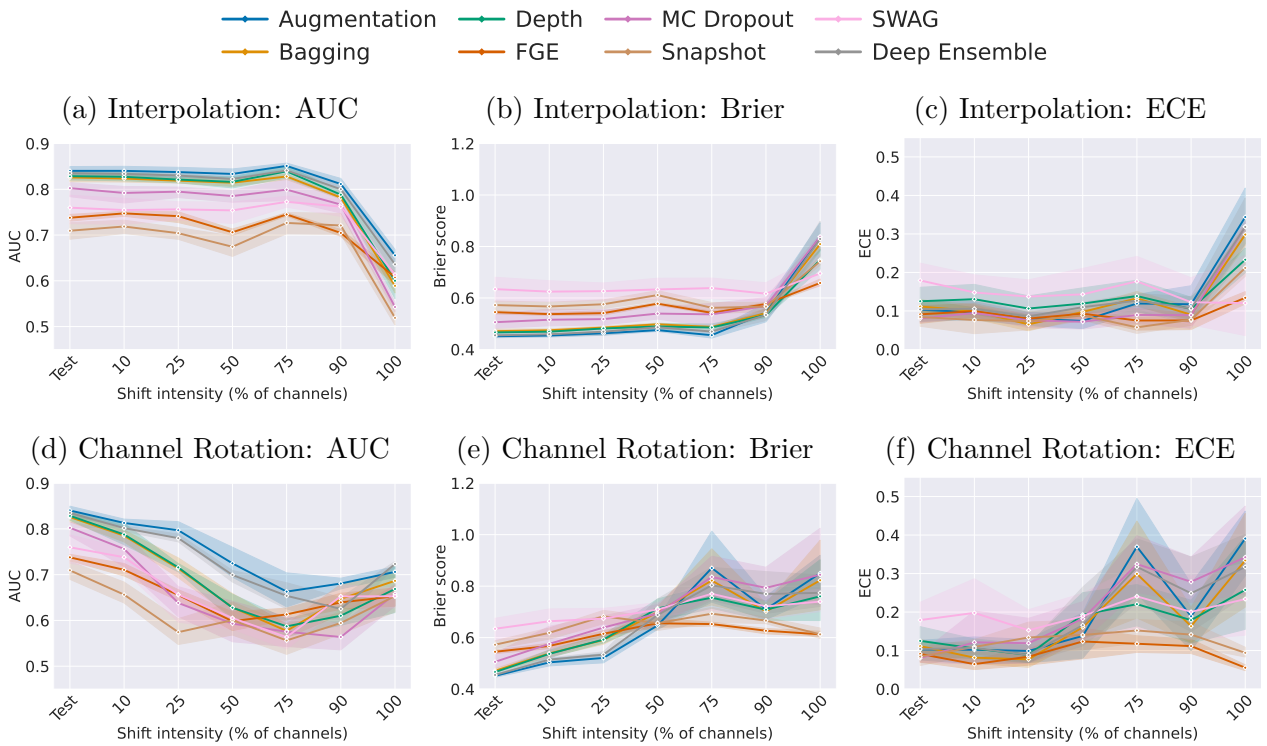

Figure 3: Illustrating ensemble performance, uncertainty and calibration under four dataset shifts. Columns plot three metrics, AUC, Brier score and ECE, against shift intensity. Rows (top to bottom) represent shift types: (1) interpolation (removing channels and interpolating), and (2) rotation (randomly rotating channel positions). Ensemble methods are labelled in the top of the figure.

4.1. In-Distribution test

Our best performing multi-run ensembles (Augmentation and Depth) achieved an AUC of 0.84±0.02 and an accuracy of 68.21%±2.75, falling short of the 0.90 AUC and 74.66% accuracy reported in the original CAUEEG paper [28]. However, the reference model relied on hybrid 1D+2D CNNs, test-time augmentation (which have shown to improve performance [53]), and over 250 million parameters. In contrast, our ensembles used lightweight architectures with only 2.5 million parameters. Other published studies report even higher EEG classification accuracies in similar classification tasks (see [54] for a review), but often on smaller datasets (10-20 times of the current). In addition, many studies use epoch-based classification, evaluating short EEG segments instead of subject-level outcomes. While this artificially increases the sample size, it can produce misleading results if predictions are not aggregated per subject, which is important and considered the correct approach in medical datasets [55]. This epoch-base approach can also introduce data leakage between training and test sets [56, 57].

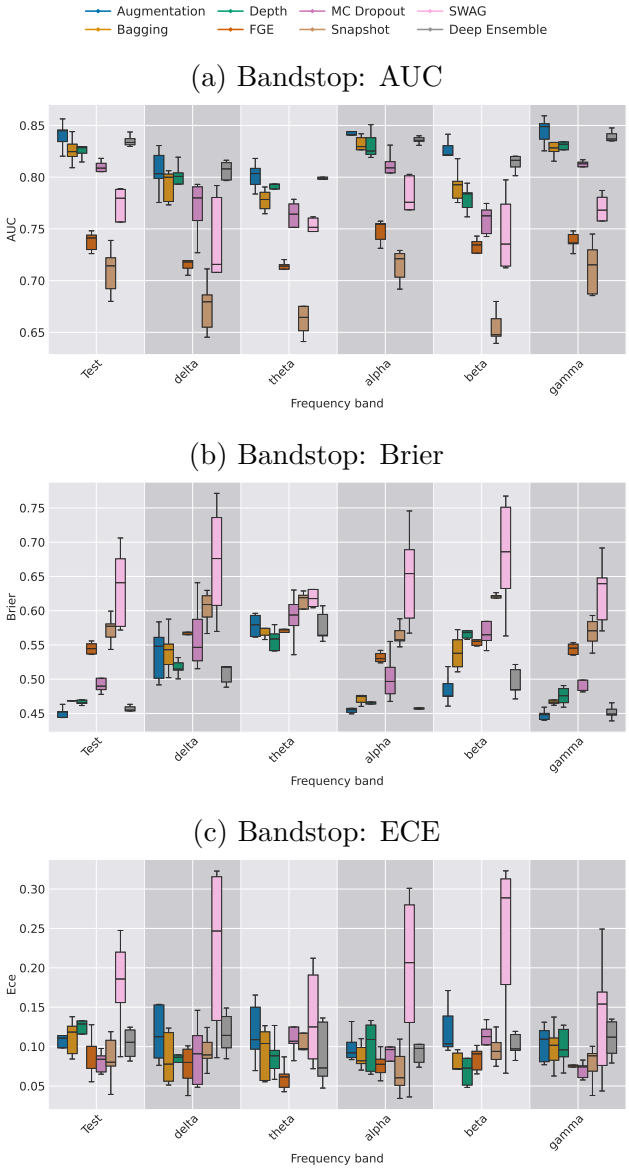

Figure 4: Performance, uncertainty and calibration of ensemble type to bandstop filtering. Rows plot, from top to bottom, (a) AUC, (b) Brier score and (c) ECE as a function of the specific frequency band removed. Ensemble methods are labelled in the top of the figure.

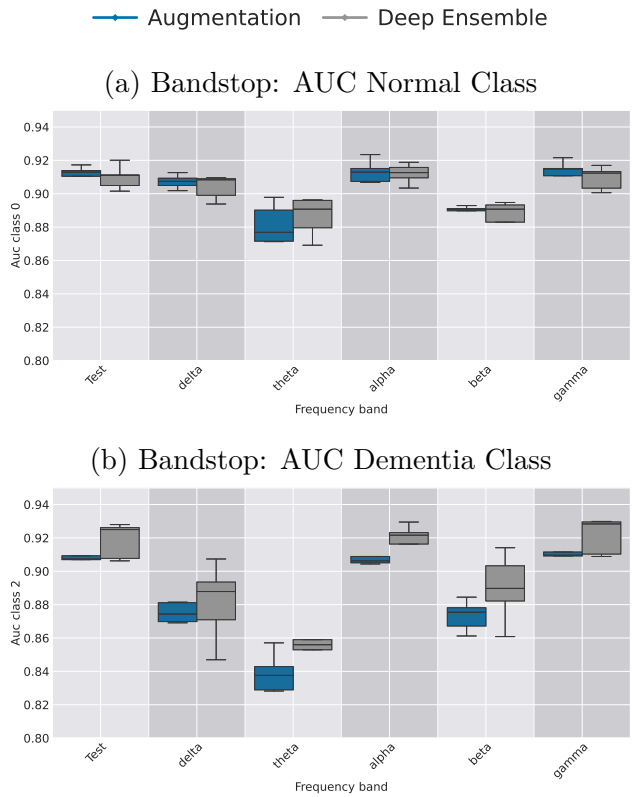

Figure 5: Per-class performance under bandstop shifts. (a) AUC for the Normal class and (b) AUC for the Dementia class, shown as a function of the frequency band removed. Results are shown for the two best-performing ensemble methods, indicated at the top of each plot. The MCI class was excluded due to unstable and uninformative results, and is therefore not displayed.

4.2. Comparing performance and uncertainty across test set and OOD

Multi-run ensembles showed consistently better performance over checkpoint- and sampling-based ensembles in both in-distribution and OOD conditions. They delivered higher predictive performance, particularly for Normal and Dementia cases (AUC $\geq$ 90), and was associated with lower uncertainty metrics. MCI remained the most difficult class, with an

AUC below 0.71, possibly reflecting its clinical heterogeneity and unstable nature [58, 59]. Checkpoint ensembles performed substantially worse across all metrics, suggesting that the limited diversity obtained from a single training trajectory is insufficient for generalisable EEG-based classification. In contrast, MCD, while based on a single model with stochastic dropout, showed intermediate performance, but with wider confidence intervals and greater variability across runs. These differences underscore the importance of ensemble diversity and architectural independence when developing models for complex clinical data.

Multi-run methods also showed higher quality of probabilistic predictions, through Brier scores, on the test set. However, the Brier scores were relatively high ($\geq$0.45), and even higher for the checkpoint and sampling methods. These values reflect limited confidence in the model outputs, most likely due to the challenging nature of MCI classification, which generally had a very poor performance, and the ecological and difficult variability in the dataset. In contrast, the ECE values were modest (0.08–0.010), suggesting moderate calibration. However, in several cases, especially for Snapshot and FGE, low ECE coincided with high Brier scores and poor accuracy, indicating underconfident but poorly performing models. This highlights the risk of over-interpreting ECE in isolation and supports the use of Brier score in addition as a more informative measure of uncertainty quality.

Under OOD conditions, differences between ensemble types became more pronounced. Multi-run ensembles kept strong performance in the Militadous cohort, with high recall for Normal cases and near-perfect precision for Dementia. MCD followed a similar trend but with greater variability, while checkpoint ensembles and SWAG degraded substantially, showing wide confidence intervals and reduced precision and recall. In the simpler single-class datasets (MPI, TDBRAIN), multi-run ensembles identified normal cases with near-perfect recall and narrow confidence intervals, confirming their robustness across settings. Age differences between the training dataset (CAUEEG; mean age 71.13) and OOD datasets (Miltiadous: 66.17, MPI: 38.26, TDBRAIN: 29.02) exist. Age was included as an input feature due to its clinical relevance and predictive value, and was regularized during training to reduce over-reliance. While prior analyses on the test set suggested that the model's sensitivity to age was limited, this is not further examined in the current paper. The strong performance on younger OOD cohorts may therefore still partly reflect demographic effects, but assessing this more directly is left for future work.

The uncertainty metrics reflected these shifts. Brier scores increased in the Militadous cohort across all methods, most notably for checkpoint- and sampling-based ensembles, indicating reduced predictive confidence. SWAG was a partial exception, showing lower average Brier scores but much wider confidence intervals, reflecting instability rather than improved reliability. Multi-run ensembles showed smaller increases in Brier score compared to other methods. In the single-class OOD datasets, Brier scores improved, most likely due to task simplicity with only the normal class which the model were generally better to classify, while ECE increased across all models. This divergence illustrates that calibration can degrade even when accuracy improves, reinforcing the importance of interpreting Brier and ECE together. Notably, the relatively low Brier scores in these simpler OOD tasks suggest that the elevated scores observed

on the test set are less a sign of model weakness and more a reflection of the in-distribution data's complexity, particularly in the poorly separated MCI class.

4.3. Clinical diagnosis

In addition to the distribution shifts explored in this study, an important source of variability lies in how diagnostic labels are assigned, both within and across datasets. This is particularly relevant when evaluating clinical conditions such as MCI. Diagnostic labels in these domains are functional, and reflect a clinical evaluation of the patient and not definitive biological markers. As such, the diagnostic labelling is subject to observer variability and local protocol differences [60, 61, 62]. These epistemic uncertainties likely contributed to the reduced model performance and elevated uncertainty observed, especially for MCI. This underscores the need for caution when interpreting model results on clinically defined labels.

4.4. Dataset shifts

Under distribution shift, we ideally expect model accuracy to decline while predictive uncertainty rises in tandem, allowing downstream systems or clinicians to recognise when predictions become unreliable. Well-calibrated models should maintain meaningful calibration even under shift, not by preserving low error, but by signalling increasing uncertainty as the input distribution diverges [13]. This behaviour enables uncertainty-aware decision making and makes predictive reliability transparent beyond the training domain.

Across the ten dataset shifts, multi-run ensembles consistently demonstrate greater robustness than checkpoint- and sampling-based methods, both in terms of predictive performance and uncertainty estimation. In most shifts, such as amplitude changes, Gaussian noise, and phase shift, performance declines gradually with increasing shift intensity, and uncertainty measures (Brier score and ECE) rise in parallel, especially for the better-performing multi-run models. This alignment suggests that their uncertainty estimates remain informative as conditions degrade. Among the sampling-based methods, MCD most closely resembles the behaviour of multi-run ensembles, showing responsive uncertainty and calibration, but with wider confidence intervals and more pronounced performance degradation. In contrast, SWAG performs inconsistently across shifts, with unstable predictions and wide variability in both performance and uncertainty, limiting its reliability. Under shifts such as timewarp, circular shift, and heavy channel rotation, all models experience sharper performance drops. Even then, Deep and Augmentation ensembles maintain comparatively higher performance. Notably, checkpoint ensembles often show flatter Brier and ECE curves even as performance deteriorates. While flat calibration metrics may appear favourable in isolation, in this context they indicate that the model fails to adjust its confidence in response to increasing error, an undesirable property under distribution shift. In contrast, multi-run ensembles generally scale their uncertainty more appropriately, with both Brier and ECE increasing alongside error. The bandstop and phase shifts revealed frequency specific vulnerabilities, especially in the theta and delta ranges, but again the relative ensemble ranking remains consistent, with multi-run models outperforming others. A closer look at the amplitude and Gaussian noise shifts reveals a distinct pattern: checkpoint ensembles

maintain stable AUC, Brier, and ECE values across the full intensity range. This apparent invariance is notable, particularly under extreme conditions where multi-run ensembles begin to degrade. However, this stability appears to be shift-specific, as checkpoint models perform less consistently under other transformations. Overall, these results highlight that multi-run ensembles are not only more accurate but also more reliable under dataset shift, an essential property for deployment in dynamic clinical environments. This aligns with previous research, which has documented the strong performance of such ensembles in different contexts [13]. While their performance declined with increasing shift severity, their uncertainty estimates scaled appropriately, indicating adaptive and informative behaviour. In contrast, checkpoint ensembles often exhibited flat uncertainty curves despite worsening predictions, suggesting a disconnect between confidence and accuracy. This lack of responsiveness may appear robust, but ultimately limits their reliability. This clearly shows that ensemble strategies vary in their response to different types of dataset shift, highlighting the importance of assessing performance and uncertainty together.

4.5. Robustness to temporal, amplitude and structural shifts

The dataset shifts explored in this study were constructed by systematically altering specific characteristics of the EEG signal, including temporal distortions, frequency-specific manipulations, amplitude scaling, and structural disruptions. While all models exhibited some vulnerability, the impact of shifts varied substantially. Examining all shifts across all models may reveal which perturbations are most disruptive and which signal properties models consistently depend on, insights that could inform both model development and EEG preprocessing strategies. Amplitude shifts led to steeper performance drops under strong amplification ($\times 8$) than under extreme reduction ($\times 0.001$), showing that over-amplified signals distort features more severely. Gaussian noise had limited impact at low levels but became damaging as the noise increased, indicating some tolerance to moderate perturbations. Timewarp and circular shift caused abrupt performance decline at the first shift level. For timewarp, this was followed by a gradual decline; for circular shift, performance remained relatively stable after the initial drop. This pattern may reflect edge effects introduced when appending the time signal to itself, and possibly the translational invariance of CNNs, which may reduce the sensitivity to the degree of temporal offset once the signal is disrupted. Slow drifts also degraded performance, highlighting the importance of removing low-frequency components, e.g., through high-pass filtering. Structural distortions had varied effects. In interpolation, performance remained stable until more than 75% of the channels were removed, which may have been due to central channels Cz and Pz being consistently preserved. Beyond that threshold, the accuracy dropped sharply. Channel rotation led to a steady decline, underscoring the importance of preserving spatial structure and consistent channel ordering between training and testing.

4.6. Frequency band importance and clinical alignment

Bandstop shifts revealed that the delta, theta, and beta bands are particularly informative for this prediction task, with their removal causing measurable performance drops. A per-

class analysis provided more detail: for the Normal class, removing either theta or beta led to a notable drop in AUC, suggesting both bands contribute to the model's ability to distinguish normal EEG. In contrast, dementia classification showed the strongest sensitivity to theta-band removal, with a more modest impact from delta and beta.

Although this study primarily focused on method development, some findings align with established EEG markers in dementia. Specifically, the model's reliance on theta, delta, and beta bands reflects well-documented spectral changes in Alzheimer's disease, where increased slow-wave activity and reduced beta power are linked to cortical disconnection and cognitive decline [63, 64, 65].

In contrast, the alpha band's contribution was minimal, a surprising finding given its well-documented importance in clinical literature. This conclusion is supported by two distinct lines of evidence: the direct removal of the alpha band via bandstop and the separate alpha peak-shift experiment both resulted in only a gradual performance decline. Although the peak-shift manipulation was less controlled, using random, bidirectional, and channel-specific changes, the consistency of its outcome with the more direct bandstop analysis lends weight to the conclusion. This limited role of alpha is likely attributable to the data originating from a mixed pre-task period (including both eyes-open and eyes-closed states) and not from a dedicated eyes-closed resting state, where alpha rhythms are most prominent. Thus, in this limited sample, the alpha band did not emerge as a significant predictive feature for the implemented models.

4.7. Limitations and future work

Our study has several limitations that provide avenues for future research. First, while the primary findings are based on a test set of 117 subjects, the robustness of our approach was further stressed against several OOD datasets. Although this represents an extensive evaluation, the conclusions still require validation on larger and more diverse multi-centre samples to establish broader generalisability. Second, our OOD evaluation focused on the Normal and Dementia classes and did not include cases of MCI. Assessing model robustness for this group is a critical future direction; however, it is crucial to recognize that this is complicated by the known clinical heterogeneity of MCI. Future work involving MCI must carefully account for variability in diagnostic criteria between datasets, which acts as a major confounding factor in the absence of consistent biological markers. These clinical inconsistencies, combined with technical differences in EEG recording hardware and preprocessing pipelines across sites, pose a significant challenge to model generalisability. Furthermore, we employed a deliberately minimal preprocessing pipeline to assess robustness in a manner that better reflects scalable deployment. While this choice may have constrained optimal model performance, it offered a test of real-world generalisability. Finally, the synthetic shifts used here for evaluation could themselves be leveraged as a data augmentation strategy during training, a promising approach to proactively enhance model robustness and clinical translation.

5. Conclusion

Multi-run ensembles, such as deep ensemble, consistently delivered the best perfor-

mance and most reliable uncertainty estimates across test, OOD, and shifted EEG data. Their ability to align uncertainty closely with changes in predictive performance under distribution shifts makes them the most promising approach for robust, uncertainty-aware EEG classification. Together, these findings underscore the value of ensemble-based uncertainty modelling as a foundation for building trustworthy and clinically deployable EEG classification systems.

Project Information and Contributions

Conflict of Interest

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

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964220. This publication reflects views of the authors and the European Commission is not responsible for any use that may be made of the information it contains.

Author Contribution

Mats Tveter: Concept, Software, Investigation, Methodology, Validation, Formal Analysis, Visualisation, Writing - Original Draft, Writing - Review & Editing **Thomas Tveitstøl:** Visualisation, Writing - Review & Editing **Christoffer Hatlestad-Hall:** Supervision, Writing - Review & Editing **Hugo L. Hammer:** Supervision, Visualisation, Writing - Review & Editing **Ira R. J. Hebold Haraldsen:** Project administration, Writing - Review & Editing, Funding acquisition, Resources

Data Availability

The dataset is accessible as open source or upon request. For further details, consult the studies: CAUEEG [28], Militadous [31] MPI [32], and TDBRAIN [33].

Code for all experiments and models can be found on GitHub‖.

References

[1] Geoffrey Hinton. "Deep Learning—A Technology With the Potential to Transform Health Care". In: *JAMA* 320.11 (Sept. 2018), pp. 1101–1102. ISSN: 0098-7484. DOI: 10.1001/jama.2018.11100. eprint: https://jamanetwork.com/journals/jama/articlepdf/2701666/jama_hinton_2018_vp_180096.pdf. URL: https://doi.org/10.1001/jama.2018.11100.

[2] Pranav Rajpurkar et al. "AI in health and medicine". In: *Nature medicine* 28.1 (2022), pp. 31–38.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[4] Khansa Rasheed et al. "Explainable, trustworthy, and ethical machine learning for healthcare: A survey". In: *Computers in Biology and Medicine* 149 (2022), p. 106043. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2022.106043.

‖ https://github.com/matstveter/eeg_dl_uncertainty

[5] Silvia Seoni et al. "Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023)". In: *Computers in Biology and Medicine* 165 (2023), p. 107441. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2023.107441. URL: https://www.sciencedirect.com/science/article/pii/S001048252300906X.

[6] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.

[7] Andreas Holzinger et al. "What do we need to build explainable AI systems for the medical domain?" In: *arXiv preprint arXiv:1712.09923* (2017).

[8] Dominik Seuß. "Bridging the Gap Between Explainable AI and Uncertainty Quantification to Enhance Trustability". In: *CoRR* abs/2105.11828 (2021). arXiv: 2105.11828. URL: https://arxiv.org/abs/2105.11828.

[9] Jakob Gawlikowski et al. "A survey of uncertainty in deep neural networks". In: *Artificial Intelligence Review* 56.Suppl 1 (2023), pp. 1513–1589.

[10] Armen Der Kiureghian and Ove Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009). Risk Acceptance and Risk Communication, pp. 105–112. ISSN: 0167-4730. DOI: https://doi.org/10.1016/j.strusafe.2008.06.020.

[11] Eyke Hüllermeier and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110 (2021), pp. 457–506. DOI: 10.1007/s10994-021-05946-3.

[12] Arsenii Ashukha et al. "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning". In: *arXiv preprint arXiv:2002.06470* (2020).

[13] Yaniv Ovadia et al. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[14] Joaquin Quiñonero-Candela et al. *Dataset shift in machine learning*. Mit Press, 2022.

[15] Dan Hendrycks and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016).

[16] Dario Amodei et al. *Concrete Problems in AI Safety*. 2016. arXiv: 1606.06565 [cs.AI]. URL: https://arxiv.org/abs/1606.06565.

[17] Juozas Vaicenavicius et al. *Evaluating model calibration in classification*. 2019. arXiv: 1902.06977 [cs.LG]. URL: https://arxiv.org/abs/1902.06977.

[18] Andrea Biasiucci, Benedetta Franceschiello, and Micah M Murray. "Electroencephalography". In: *Current Biology* 29.3 (2019), R80–R85.

[19] Yannick Roy et al. "Deep learning-based electroencephalography analysis: a systematic review". In: *Journal of neural engineering* 16.5 (2019), p. 051001.

[20] Jesper Tveit et al. "Automated Interpretation of Clinical Electroencephalograms Using Artificial Intelligence". In: *JAMA Neurology* 80.8 (Aug. 2023), pp. 805–

812. ISSN: 2168-6149. DOI: 10.1001/jamaneurol.2023.1645. eprint: https://jamanetwork.com/journals/jamaneurology/articlepdf/2806244/jamaneurology_tveit_2023_oi_230036_1691786999.53858.pdf. URL: https://doi.org/10.1001/jamaneurol.2023.1645.

[21] Xiyuan Jin et al. "Uncertainty-Aware Denoising Network for Artifact Removal in EEG Signals". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), pp. 4470–4480. DOI: 10.1109/TNSRE.2023.3330963.

[22] Daily Milanés-Hermosilla et al. "Monte Carlo Dropout for Uncertainty Estimation and Motor Imagery Classification". In: *Sensors* 21.21 (2021). ISSN: 1424-8220. DOI: 10.3390/s21217241. URL: https://www.mdpi.com/1424-8220/21/21/7241.

[23] Sheng Wong et al. "Estimating Patient-Level Uncertainty in Seizure Detection Using Group-Specific Out-of-Distribution Detection Technique". In: *Sensors* 23.20 (2023). ISSN: 1424-8220. DOI: 10.3390/s23208375. URL: https://www.mdpi.com/1424-8220/23/20/8375.

[24] Huy Phan et al. "SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification". In: *IEEE Transactions on Biomedical Engineering* 69.8 (2022), pp. 2456–2467. DOI: 10.1109/TBME.2022.3147187.

[25] Luigi Fiorillo, Paolo Favaro, and Francesca Dalia Faraci. "DeepSleepNet-Lite: A Simplified Automatic Sleep Stage Scoring Model With Uncertainty Estimates". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), pp. 2076–2085. DOI: 10.1109/TNSRE.2021.3117970.

[26] Tiehang Duan et al. "UNCER: A Framework for Uncertainty Estimation and Reduction in Neural Decoding of EEG Signals". In: *Neurocomputing* 538 (Apr. 2023). DOI: 10.1016/j.neucom.2023.03.071.

[27] Mats Tveter et al. "Advancing EEG prediction with deep learning and uncertainty estimation". In: *Brain Informatics* 11.1 (2024), pp. 1–12.

[28] Min-jae Kim, Young Chul Youn, and Joonki Paik. "Deep learning-based EEG analysis to classify normal, mild cognitive impairment, and dementia: Algorithms and dataset". In: *NeuroImage* 272 (2023), p. 120054. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2023.120054.

[29] Mainak Jas et al. "Autoreject: Automated artifact rejection for MEG and EEG data". In: *NeuroImage* 159 (2017), pp. 417–429.

[30] Martin Prince et al. "The global prevalence of dementia: a systematic review and metaanalysis". In: *Alzheimer's & dementia* 9.1 (2013), pp. 63–75.

[31] Andreas Miltiadous et al. "A dataset of EEG recordings from: Alzheimer's disease, Frontotemporal dementia and Healthy subjects". In: *OpenNeuro* 1 (2023), p. 88.

[32] Anahit Babayan et al. "A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults". In: *Scientific Data* 6.1 (2019), pp. 1–21.

[33] Hanneke Van Dijk et al. "The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database". In: *Scientific data* 9.1 (2022), p. 333.

[34] Cédric Rommel et al. "Data augmentation for learning predictive models on EEG: a systematic comparison". In: *Journal of Neural Engineering* 19.6 (2022), p. 066020.

[35] Matti S Hämäläinen and Risto J Ilmoniemi. "Interpreting magnetic fields of the brain: minimum norm estimates". In: *Medical & biological engineering & computing* 32 (1994), pp. 35–42.

[36] Hassan Ismail Fawaz et al. "InceptionTime: Finding AlexNet for time series classification". In: *Data Mining and Knowledge Discovery* 34.6 (Nov. 2020), pp. 1936–1962. ISSN: 1573-756X. DOI: 10.1007/s10618-020-00710-y. URL: https://doi.org/10.1007/s10618-020-00710-y.

[37] Thomas Tveitstøl et al. "Introducing Region Based Pooling for handling a varied number of EEG channels for deep learning models". In: *Frontiers in Neuroinformatics* 17 (2024). ISSN: 1662-5196. DOI: 10.3389/fninf.2023.1272791.

[38] Alireza Rafiei et al. "Automated Detection of Major Depressive Disorder With EEG Signals: A Time Series Classification Using Deep Learning". In: *IEEE Access* 10 (2022), pp. 73804–73817. DOI: 10.1109/ACCESS.2022.3190502.

[39] Thomas Tveitstøl et al. "Assessing the robustness of deep learning based brain age prediction models across multiple EEG datasets". In: *bioRxiv* (2025). DOI: 10.1101/2025.05.20.655022. eprint: https://www.biorxiv.org/content/early/2025/05/21/2025.05.20.655022.full.pdf. URL: https://www.biorxiv.org/content/early/2025/05/21/2025.05.20.655022.

[40] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[41] Takuya Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[42] James Bergstra et al. "Algorithms for Hyper-Parameter Optimization". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[43] Chuan Guo et al. "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[44] L. Breiman. "Bagging Predictors". In: *Machine Learning* 24 (1996), pp. 123–140. URL: https://api.semanticscholar.org/CorpusID:47328136.

[45] Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[46] Gao Huang et al. "Snapshot ensembles: Train 1, get m for free". In: *arXiv preprint arXiv:1704.00109* (2017).

[47] Ilya Loshchilov and Frank Hutter. "SGDR: Stochastic Gradient Descent with Restarts". In: *CoRR* abs/1608.03983 (2016). arXiv: `1608.03983`. URL: `http://arxiv.org/abs/1608.03983`.

[48] Wesley J Maddox et al. "A simple baseline for bayesian uncertainty in deep learning". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13153–13164.

[49] Pavel Izmailov et al. "Averaging weights leads to wider optima and better generalization". In: *arXiv preprint arXiv:1803.05407* (2018).

[50] Timur Garipov et al. "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs". In: *Advances in Neural Information Processing Systems*. 2018.

[51] Andrew P Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[52] Glenn W Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[53] Alexander Lyzhov et al. "Greedy policy search: A simple baseline for learnable test-time augmentation". In: *Conference on uncertainty in artificial intelligence*. PMLR. 2020, pp. 1308–1317.

[54] Madhav Acharya et al. "Deep learning techniques for automated Alzheimer's and mild cognitive impairment disease using EEG signals: A comprehensive review of the last decade (2013-2024)". In: *Computer Methods and Programs in Biomedicine* (2024), p. 108506.

[55] Ilias Tougui, Abdelilah Jilbab, and Jamal El Mhamdi. "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications". In: *Healthcare informatics research* 27.3 (2021), pp. 189–199.

[56] Geoffrey Brookshire et al. "Data leakage in deep learning studies of translational EEG". In: *Frontiers in Neuroscience* 18 (2024), p. 1373515.

[57] Hyung-Tak Lee et al. "Risk of data leakage in estimating the diagnostic performance of a deep-learning-based computer-aided system for psychiatric disorders". In: *Scientific Reports* 13.1 (2023), p. 16633.

[58] Ronald C Petersen et al. "Mild cognitive impairment: a concept in evolution". In: *Journal of internal medicine* 275.3 (2014), pp. 214–228.

[59] Simone Salemme et al. "The prognosis of mild cognitive impairment: A systematic review and meta-analysis". In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 17.1 (2025), e70074.

[60] Federico Cabitza, Davide Ciucci, and Raffaele Rasoini. "A Giant with Feet of Clay: On the Validity of the Data that Feed Machine Learning in Medicine". In: *Organizing for the Digital World*. Ed. by Federico Cabitza, Carlo Batini, and Massimo Magni. Cham: Springer International Publishing, 2019, pp. 121–136.

[61] Anna van Oosterzee. "AI and mental health: evaluating supervised machine learning models trained on diagnostic

classifications". In: *AI & SOCIETY* (2024), pp. 1–10.

[62] Zhe Sage Chen et al. "Modern views of machine learning for precision psychiatry". In: *Patterns* 3.11 (2022).

[63] Jaeseung Jeong. "EEG dynamics in patients with Alzheimer's disease". In: *Clinical Neurophysiology* 115.7 (2004), pp. 1490–1505. ISSN: 1388-2457. DOI: https://doi.org/10.1016/j.clinph.2004.01.001. URL: https://www.sciencedirect.com/science/article/pii/S138824570400015X.

[64] Claudio Babiloni et al. "Measures of resting state EEG rhythms for clinical trials in Alzheimer's disease: Recommendations of an expert panel". In: *Alzheimer's & Dementia* 17.9 (2021), pp. 1528–1553. DOI: https://doi.org/10.1002/alz.12311. eprint: https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/alz.12311. URL: https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.12311.

[65] P.M. Rossini et al. "Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis. Report from the IFCN-sponsored panel of experts". In: *Clinical Neurophysiology* 131.6 (2020), pp. 1287–1310. ISSN: 1388-2457. DOI: https://doi.org/10.1016/j.clinph.2020.03.003. URL: https://www.sciencedirect.com/science/article/pii/S1388245720300870.