

CSCI 409 Senior Project II

Final Project Report, Spring 2026

Title: A Subject-Independent Model for Motor Imagery Brain-Computer Interfaces, with Extensions to Emotion Recognition and Seizure Detection

Team Members: Diar Duskenov, Ilyas Imankulov, Asset Bissenbayev

Advisor: Min-Ho Lee

Department: School of Engineering and Digital Sciences, Nazarbayev University

1. Executive Summary

MI BCIs that use electroencephalography (EEG) as a source of information allow controlling external devices through mental representations of hand actions. However, the biggest barrier to the practical implementation of MI BCIs beyond laboratory conditions is the difference between subject-dependent (SD) performance and subject-independent (SI) performance. When an EEG-based model is trained on one subject, it achieves 75 to 85 percent accuracy for four-class MI. Yet, if the same model is applied to another subject without calibration information, it attains only 55 to 65 percent accuracy. Closing this gap is the problem our project targets.

During Fall 2025 we implemented a PyTorch pipeline with two baselines, EEGNet and a Compact Convolutional Transformer (EEGCCT), and a knowledge-distillation (KD) framework with a CORAL domain-alignment loss and spectral transfer augmentation. All Fall evaluation used leave-one-subject-out (LOSO) on BCI Competition IV-2a with binary left-versus-right decoding. The best Fall configuration reached 75.9 percent mean LOSO accuracy with EEGNet.

In Spring 2026 the project changed direction in three ways. First, we switched the primary backbone from EEGNet to EEGNeX. EEGNeX is designed to train directly on raw EEG, so heavy fixed preprocessing is no longer a hard requirement of the pipeline. On top of EEGNeX we built two new variants. ImprovedEEGNeX is a wide, multi-branch model that combines a learnable filter bank, multi-scale temporal convolutions, SE and CBAM attention, a temporal gate, self-attention, a graph-attention channel branch, domain-adaptation losses, and knowledge distillation from a spectral teacher. EEGNeX_SI_v2 is a lean variant near 40k parameters that replaces the

depthwise BatchNorm in the spatial block with GroupNorm, which removes a concrete source of subject leakage. Second, we extended the same framework beyond MI to emotion recognition on the EAV (EEG-Audio-Video) dataset (led by Ilyas) and seizure detection on the Siena Scalp EEG dataset (led by Asset), so the pipeline now supports three EEG tasks with a shared backbone. Third, we built a Streamlit toolbox that runs every training mode (SD, SI, K-Fold) with live metrics, sweep mode, per-subject breakdowns, confusion matrices, and ROC plots, all written out as reproducible JSON summaries.

The main MI results: evaluated on the official BCI Competition IV-2a evaluation set, ImprovedEEGNeX reaches 72.0 percent mean LOSO accuracy with Cohen kappa 0.63 and AUC-ROC 0.87, while the lean EEGNeX_SI_v2 reaches 69.1 percent with kappa 0.59. In subject-dependent mode, ImprovedEEGNeX reaches 79.2 percent (kappa 0.72), the highest SD result across all four models. In comparison to the Fall, where we reached 75.9 percent binary LOSO with EEGNet, the Spring numbers are four-class and evaluated on the official held-out evaluation set with released ground-truth labels. The other major advantage is that ImprovedEEGNeX has better run-to-run stability than Fall – while our network achieved 72.0 ± 0.8 percent on the official evaluation dataset when tested across five different random seeds, Fall's peak (75.9 percent binary EEGNet) was obtained from one lucky run with considerable variance between runs. In this semester, we explicitly chose to focus on reliable and consistent performance rather than an unpredictable and lucky peak. Each reported run passed a 1000-iteration permutation test with label shuffling, where the p-value is below 0.001. Our ablation study shows that Euclidean Alignment, knowledge distillation, and filter-bank front-end are the core features.

This project includes all aspects of designing, developing, and evaluating a computing solution – we designed two new neural networks, implemented them end-to-end with a toolbox and a 12-metric evaluation harness, and evaluated them using three protocols on an open benchmark. The emotion recognition branch and seizure detection branch share the same infrastructure, waiting for numbers to be added.

2. Introduction

2.1 Problem

A motor imagery BCI turns an imagined hand movement into a digital command. The recordings are performed using a scalp EEG cap, meaning that the signals are small (microvolt-level), noisy (muscle/line noise), and individualized depending on their anatomical differences (head and electrodes placement). No matter how sophisticated the

preprocessing is, once trained on someone’s EEG data, a machine learning model will be unable to perform on another person’s data until re-calibration. Re-calibrating a BCI takes around twenty-thirty minutes in average. This is just one of many reasons that BCIs for stroke, ALS, and spinal cord injury patients are not usable on a daily basis.

2.2 Motivation

The motivation behind this project is multi-faceted, as follows. From the clinical perspective, reducing calibration time might increase usability among users who have trouble sitting through such a procedure. Secondly, from the technical perspective, MI EEG serves well as a benchmarking domain adaptation experiment, as each participant represents a separate and distinct domain (labeled). Finally, from the academic point of view, BCI Competition IV-2a serves as the most commonly cited MI EEG benchmark, and thus improving it is quantifiable in the context of literature.

2.3 Scope and expansion

Initially, our scope in Fall 2025 included MI only. For Spring 2026, we worked on expanding our scope with our advisor to include additional tasks, all of which will have the same pipeline and architectural design:

1. Motor Imagery Classification on BCI Competition IV-2a (main goal, Diar)
2. Emotion Recognition on a Public EEG Emotion Dataset (Ilyas)
3. Seizure detection on a Public Scalp EEG Seizure Dataset (Asset)

2.4 Proposed solution

The proposed solution consists of three components: two architectures and one infrastructure-wise:

1. **ImprovedEEGNeX** – an updated version of EEGNeX consisting of trainable band filter, MS-TCNs, SE/CBAM block, TG, MHSA, Graph-Attention on EEG channels and spectral knowledge distillation of a six-band model, optionally with CORAL and DANN domain adaptation heads.
2. **EEGNeX_SI_v2** – a lighter version of EEGNeX which switches out the spatial block batch norm layer for Group Norm without any subject dependent statistics for improved SI capabilities.
3. Streamlit toolbox with an evaluation harness of 12-metrics (Accuracy, Balanced Accuracy, Cohen’s Kappa, MCC, Macro and Weighted F1 score, Macro Precision and Recall, Log Loss, AUC-ROC, Permutation p-value and Chance level) supporting SD, SI, and K-fold training and JSON outputs.

2.5 Contributions

1. EEGNeX – a raw EEG architecture (backbone) replacing a fixed preprocessing pipeline used in Fall.
2. New architectures: ImprovedEEGNeX (wider version) and EEGNeX_SI_v2 (lighter version).
3. Streamlit toolkit for the interactive experience in training and evaluating.
4. Common architecture for motor imagery, emotion recognition and seizures tasks with unified architecture.
5. Evaluation harness containing 12 metrics and permutation test of 1000 repetitions.

2.6 Report organization

The list of constraints is available in Section 3. Works related to EEG deep learning, SI transfer and cross-tasks EEG will be described in Section 4. Description of the system comes in Section 5. Story of project development in two semesters will follow in Section 6. Evaluation method and results are provided in Section 7. Future work and references will be in Sections 8 and 9 respectively.

3. Identification of Constraints

3.1 Ethical

EEG data can be considered as private data, thus all EEG data should be anonymized, all processing and actions should minimize the chances of privacy violations. Additionally, end users who would use BCI with provided models may over-trust its output, therefore deployed versions must have clear confidence signaling and a no decision state.

3.2 Legal and data-protection

EEG datasets that used in this study are public thus only used for research purposes with attribution of original authors. Moreover, these raw data will not be republished, and final deliverable is only model definitions, processing pipeline and so on. It was also done in case this framework will be used on real patient data that fall under different data and privacy protection laws, calibration of the medical devices, etc.

3.3 Economic

As both models are pure CNN based and contain a small number of parameters, precisely ImprovedEEGNeX under 200k parameters and EEGNeX_SI_v2 near 40k parameters, they both can be run on consumer grade GPU workstation. In addition, to reduce computational cost, early stopping is also enabled to stop if the model is not improving over a specific number of epochs.

3.4 Environmental

The main environmental cost is GPU energy during training that was reduced again by using the early stopping. In addition, all the parameters are saved in the config file, so there is no need to rerun the model to get the same result in case of forgotten used settings.

3.5 Manufacturability

As the only deliverable is software, precisely Python script, it is more about reproducibility. However, in order to replicate all required libraries, a python version can be installed from requirements.txt, also we use configs and fixed seeds to be able to reproduce the same or close results.

3.6 Sustainability and social responsibility

The main and long term goal is to build assistive BCI that will not require any per user calibration as it is the main bottleneck of current BCIs. Thus the provided framework is open source within the university so future teams could improve current setup, test their own models, or build BCI around provided models.

3.7 Health and safety

There is no risk to health and safety as these models were tested offline only on public datasets, models weren't used on BCI at any stage.

3.8 Trade-offs and limitations

The main trade-off is between architectural capacity and subject-independent generalization. While ImprovedEEGNeX excels in subject dependent scenarios it is still sensitive to subject leakage if domain - adaptation components are disabled. On the other hand, EEGNeX_SI_v2 contains fewer parameters and stronger SI behavior but has lower SD accuracy.

4. Background and Related Work

4.1 Non-invasive BCI and BCI Competition IV-2a

The main dataset is BCI Competition IV-2a [20] which is the most widely used benchmark for four-class motor imagery. It contains nine subjects (A01 to A09), 22 EEG channels at 250 Hz, and 2 separate sets of 288 trials per subject across left hand, right hand, feet, and tongue. Non-invasive BCIs and related problems are surveyed by Edelman et al. [1], who highlighted that one of main challenges is calibration and subject variability.

4.2 Deep models for EEG

EEGNet [12] introduced a compact depthwise-separable convolutional block that remains the reference baseline. DeepConvNet and ShallowConvNet [13] were the first end-to-end EEG CNNs and are still used as sanity checks. EEGNeX [14] extends EEGNet with a second temporal convolution and dilated convolutions, and is designed to work directly on raw EEG, which reduces dependence on fixed preprocessing. Compact convolutional transformers applied to subject-independent MI were reported by Keutayeva et al. [3], showing that attention-based tokenizers can match CNN baselines. SVM-enhanced attention for MI was reported by Otarbay and Kyzyrkanov [8]. Kwon et al. [4] studied subject-independent BCIs with deep CNNs and provided an early benchmark for the SD-SI gap.

4.3 Subject-independent and domain-adaptation techniques

Subject independence on EEG typically needs data-level alignment or loss-level alignment, or both. Euclidean Alignment [7] pre-whitens each subject's channel covariance to identity, which moves every subject into a common reference frame and is the single largest lever in our ablation. CORAL [22] aligns second-order feature statistics between source and target. DANN [21] adds a gradient-reversal domain classifier that pushes the encoder to produce subject-invariant features. Adaptive BatchNorm [7] re-estimates BN statistics from target data without gradient updates. Knowledge distillation [24] transfers softened probabilities from a teacher model to a student, and the spectral-teacher variant reported by Zheng et al. [7] uses a phase-invariant teacher to regularize the student toward domain-generalized features. Hybrid neural networks that combine filter banks, GAN-based augmentation, and attention were evaluated by Zhang et al. [6].

4.4 Graph and attention mechanisms

EEG channels have a natural spatial graph structure. Graph Attention Networks [17] operate on the channel-correlation adjacency matrix, which lets the model learn subject-level topology. SE [15] and CBAM [16] are standard channel and spatial attention blocks and are cheap to plug into convolutional backbones. Lu et al. [5] applied a dual-branch CNN with graph attention and adversarial domain adaptation to cross-subject EEG emotion recognition, which motivated the GAT branch in our ImprovedEEGNeX.

4.5 Cross-task EEG (emotion and seizure)

The same preprocessing and backbone families used for MI are also standard in EEG emotion recognition and seizure detection. For emotion recognition, we use the EAV (EEG-Audio-Video) dataset [25], which contains EEG recordings paired with audio and video modalities for emotion recognition in conversational contexts. For seizure detection, the Siena Scalp EEG database [26] is used; event-level metrics (sensitivity, false-alarm rate per hour, detection latency) are typical for this task. Ren et al. [9] report

cross-task transfer learning between error-related potentials and other EEG tasks, which supports the idea of reusing a shared backbone.

4.6 Gap and justification

No published architecture we reviewed combines a learnable filter-bank front end, multi-scale temporal convolutions, SE and CBAM, a temporal gate, multi-head self-attention, a graph-attention channel branch, and knowledge distillation in a single evaluated system with a rigorous 12-metric harness. ImprovedEEGNeX fills this gap on the additive side. EEGNeX_SI_v2 fills a second gap on the subtractive side by showing that removing BatchNorm from the spatial block is a larger SI win than adding most individual components.

5. Project Approach

5.1 Data and preprocessing

The primary dataset is BCI Competition IV-2a. Each subject contributes 288 four-second trials at 250 Hz across 22 EEG channels, which gives an input tensor of shape (22, 1000) per trial and four class labels.

The Fall 2025 preprocessing used a 4 to 38 Hz bandpass, ICA-based artifact rejection, epoching, and channel-wise z-score standardization, saved as a dense .npz per subject. In Spring 2026 we moved to EEGNeX as the primary backbone. EEGNeX is designed to consume raw EEG and absorbs most of the temporal filtering inside its learned convolutional blocks, which means the fixed preprocessing stack is no longer load-bearing. We kept an optional bandpass and z-score step for comparison with Fall baselines, but all Spring results use the raw-EEG path with only standardization.

The emotion recognition branch uses the EAV (EEG-Audio-Video) dataset [25], which provides EEG recordings in conversational emotion contexts. The seizure detection branch uses the Siena Scalp EEG database [26]. Both branches reuse the same loader abstraction: each plugs in its own dataset adapter and label scheme, but the training loop, evaluation harness, and Streamlit toolbox are shared.

5.2 Baselines

We keep three baselines: EEGNet, EEGCCT, and EEGNeX_8_32. EEGNet is the faithful PyTorch port we used in Fall. EEGCCT is the compact convolutional transformer used in Fall. EEGNeX_8_32 is the Keras-to-PyTorch port of the original EEGNeX, with a second temporal convolution and dilated convolutions. These three cover the CNN, transformer, and raw-EEG CNN families.

5.3 ImprovedEEGNeX

ImprovedEEGNeX is built on top of EEGNeX with the following components, each motivated by one prior paper:

- **Learnable five-band filter bank** (delta, theta, alpha, beta, gamma) with softmax-weighted fusion and a one-by-one conv, following the filter-bank line [11].
- **Multi-scale temporal convolution** with three parallel branches (kernel 5, 10, 25) and per-branch BatchNorm, which captures MI rhythms at different time scales.
- **Depthwise spatial convolution** across the 22 channels, followed by SE and CBAM channel and spatial attention [15, 16].
- **NMA temporal gate**: a sigmoid gate on the flattened temporal features that suppresses low-information windows.
- **Multi-head self-attention** over tokenized temporal windows, with four heads and a token dimension of 64.
- **Graph-Attention channel branch**: a small GAT [17] over a channel-correlation adjacency matrix, fused with the CNN branch through learned weights.
- **Domain-adaptation losses**: Euclidean Alignment [7] before pooling, Adaptive BatchNorm at test time, optional CORAL [22] and DANN [21] heads.
- **Knowledge distillation** using a six-band spectral teacher [7] using KL + feature-match loss, motivated by label smoothing 0.1 and temperature 4.
- **Pretraining on the entire population** [7] excluding target individuals prior to fine-tuning, which helps SI initialization.

Default number of parameters used is approximately 150k for the baseline setup. Every module uses a corresponding flag in the config, allowing arbitrary ablation in Section 7.

5.4 EEGNeX_SI_v2

EEGNeX_SI_v2 is a minimal redesign of EEGNeX that targets subject-independent transfer. The key change is in the spatial block: the depthwise BatchNorm is replaced with GroupNorm (eight groups across 32 channels). The motivation: BN running statistics on the spatial block learn each training subject's channel-wise mean and variance, and these are re-used at test time on a different subject, which biases the spatial features. GroupNorm normalizes per sample, so it cannot memorize subject statistics. We chose GroupNorm over InstanceNorm because InstanceNorm removes the constant component that encodes the ERD/ERS power asymmetry between the μ (8 to 12 Hz) and beta (13 to 30 Hz) bands. Our model retains a total parameter count of 40k, maintains BN for temporal-only modules (rhythms being universal across subjects), and uses one SE block in the dilated temporal pathway. Max-norm constraints follow EEGNet and EEGNeX conventions.

5.5 Streamlit toolbox

We wrote a Streamlit application that turns the trainer into an interactive tool. The user picks the preprocessed dataset folder, selects the model (EEGNet, EEGNeX_8_32, EEGCCT, ImprovedEEGNeX, EEGNeX_SI_v2), chooses the training mode (SD, SI, or K-Fold), and sets hyperparameters (epochs, learning rate, weight decay, dropout, KD flags, domain-adaptation flags). The app shows live training and validation curves, per-fold metrics, per-subject accuracy for SI, confusion matrices, ROC curves, and a sweep mode that iterates over any subset of the gated architectural flags. Each run creates a timestamped folder with a JSON summary file, PNGs for all plots, and a dump of the config. This process improved the previous script editing and rerunning approach that took tens of minutes per iteration and reduced the time per experiment to a few minutes.

5.6 Evaluation harness

For each fold, the harness calculates twelve evaluation measures: accuracy, balanced accuracy, baseline accuracy, Cohen’s kappa, Matthews’ correlation coefficient, macro and weighted F1 scores, macro precision and recall, cross-entropy loss, AUC ROC (one-versus-rest), and a permutation p-value calculated based on one thousand permutations of labels to determine when a random classifier has a similar performance to the model in terms of accuracy. Each measure is reported separately per fold in the JSON summary file.

The reproducibility is ensured using a global seed, deterministic cuDNN kernel, and seeded DataLoader workers. Two consecutive runs with the same config generate identical output files across the entire pipeline on a single machine.

5.7 Cross-task extension

Apart from motor imagery classification, we explored two more applications with our framework: emotion recognition (Ilyas) and seizure detection (Asset). For emotion recognition, the model is applied to the EAV (EEG-Audio-Video) dataset, which contains EEG recordings captured during conversational scenarios and labeled by emotional state. For seizure detection, the model is applied to the Siena Scalp EEG database, a clinical dataset of real patient recordings; unlike the trial-level accuracy used for MI, seizure detection is evaluated on event-level metrics such as sensitivity, false alarms per hour, and detection latency. Both tasks reuse the same training pipeline, evaluation harness, and Streamlit toolbox as the MI work, with each task providing its own dataset adapter and label scheme. Results for both extensions are reported in Section 7.

5.8 Third-party components

We use PyTorch for models and training, NumPy and scikit-learn for data and metrics, MNE-Python for raw EEG loading and ICA, Streamlit for the toolbox, and Matplotlib plus seaborn for plots. We also use `torch.optim.swa_utils` for optional Stochastic

Weight Averaging. Everything model-specific (architectures, losses, augmentation, training loop, harness, Streamlit tabs, sweep mode, seeding) is implemented by the team.

5.9 Team roles

- **Diar Duskenov**: MI modeling, ImprovedEEGNeX and EEGNeX_SI_v2 design, PyTorch modules, unit tests for forward-shape correctness.
 - **Ilyas Imankulov**: Emotion recognition branch, preprocessing and evaluation harness, Streamlit toolbox.
 - **Asset Bissenbayev**: Seizure detection branch, dataset analysis, event-level evaluation, preprocessing for scalp seizure EEG.
 - **Advisor Min-Ho Lee**: Weekly methodology review and guidance on subject-independent design choices.
-

6. Project Execution

6.1 Fall 2025: baselines and first improvements

We began by porting EEGNet and EEGNeX to PyTorch from the original Keras code, rather than using a third-party port, because existing ports silently omit the second temporal convolution and the dilated-conv block. This paid off later, because every number we report is comparable to the original papers. On top of these baselines we built EEGCCT and a knowledge-distillation framework with a phase-only teacher and a CORAL alignment loss. Spectral transfer augmentation was added to improve cross-subject stability.

The Fall evaluation protocol was binary left-versus-right LOSO on BCI Competition IV-2a. The baseline EEGCCT reached 72.9 percent mean LOSO accuracy with moderate per-subject variance. EEGNet reached 75.9 percent and was the strongest Fall model. Knowledge distillation improved confidence calibration for both EEGCCT and EEGNet but did not move the overall accuracy headline.

The Fall evaluation used the BCI Competition IV-2a training sessions only, because the competition evaluation-set ground-truth labels had not yet been publicly released. The official evaluation set (session 2 per subject) became available in Spring 2026, enabling the canonical cross-session benchmark protocol for the first time in our pipeline.

We also observed a first SI attempt with EEGNeX collapsing to near chance. A careful read of the training logs showed that BatchNorm running statistics were learning subject-specific spatial patterns. This finding drove the design of EEGNeX_SI_v2 in Spring.

6.2 Transition decision: Fall to Spring

Two decisions were made at the start of Spring. First, we switched the primary backbone from EEGNet to EEGNeX. EEGNeX works directly on raw EEG, which removes the dependency of every result on a fixed preprocessing stack and gives us a cleaner ablation axis. Second, we deprioritized EEGCCT for the main track after repeated LOSO runs showed high variance between runs with the same seed. EEGCCT is kept as a baseline in the evaluation tables, but the Spring architectural work is all on top of EEGNeX.

6.3 Spring 2026: building in layers

Rather than wire up every component of ImprovedEEGNeX at once, we added them in layers and ran an ablation after each one. The order was: filter bank, multi-scale temporal, attention stack (SE plus CBAM), temporal gate, multi-head self-attention, GAT branch, domain-adaptation losses, and knowledge distillation from the spectral teacher. Each component is gated by a flag in the config, which made the ablation in Section 7 a direct read-out of the toggles.

EEGNeX_SI_v2 came out of the BatchNorm investigation. We drew the EEGNeX block diagram on paper, marked every normalization layer, and concluded which ones could be memorizing subject statistics. The spatial-block BN was the obvious candidate, because the spatial filter is where the subject's skull geometry bites. Replacing that single BN with GroupNorm was the smallest change that moved the SI number.

6.4 Things that went wrong

Following problems cost time.

First of all, knowledge distillation did not converge. Two bugs were overlaid: the teacher was being trained in `train()` mode, letting its BN statistics drift during KD, and the student KD trainer did not recover the best checkpoint after early stopping. Although both fixes are moderate, they required a week to find out due to the symptom being slow divergence as opposed to a crash.

Second, `copy.deepcopy(state_dict())` silently made shallow copies in CUDA on Windows. This lost each model 20 to 40 percentage points of accuracy a complete week before we picked it up. The solution is to explicitly `.detach().clone()` all the tensors in the state dict. We now consider `copy.deepcopy` on a CUDA state dict as a prohibited pattern in the codebase.

Third, one of the sweep configs was left with a `weight_decay = 0.02` value and collapsed ImprovedEEGNeX to about 35 percent accuracy in a few runs. We now write the safe range of weight decay in a best-configs document in the repository.

6.5 Things that went right

The most successful engineering investment of the semester was the Streamlit toolbox. Shifting training and ablation to an interactive application reduced the iteration time,

previously about half an hour per experiment (edit script, re-run, search logs), to just a few minutes, and sweep mode was now a practical option.

The 12-metric harness revealed issues that were hidden by raw accuracy. A fixed initial setup initially gave reasonable accuracy but a Cohen kappa of about zero, which actually was majority-class prediction on an unbalanced debug split. Subsequently, the permutation test established that such an arrangement could not have been statistically different from that of chance. This would not have been caught so accurately alone.

Towards the end of the semester, the discipline of reproducibility came in handy. When seeding the deterministic cuDNN, DataLoader worker, and seeding the seed, back-to-back runs were bit-identical. This rendered the ablation table reliable.

6.6 Teamwork

The codebase was divided into clean modules which allowed the three of us to work at the same time without merging code: MI models and training (Diar), emotion recognition branch and evaluation harness and Streamlit toolbox (Ilyas), seizure detection branch and event-level evaluation (Asset). The methodology was maintained by weekly meetings with the advisor particularly on SI evaluation protocol and statistical rigor. We were advised into using Cohen kappa and the permutation test at the beginning of Spring, which changed some of our internal rankings.

6.7 Planning artifacts

We kept our versions in GitHub and reviewed our code, and kept our report and shared documentation in Google Drive. Each member created temporary branches on which he opened independent features and merged with pull requests and executed the Streamlit toolbox against the merged main and then reported numbers. The safe hyperparameter ranges and foot-guns (such as the weight-decay range and the CORAL-versus-DANN interference) are documented in a best-configs file in the repository, such that they are not rediscovered in the future by other teammates.

7. Evaluation

7.1 Evaluation protocol

Every model is evaluated under 3 modes:

- **Subject-Dependent (SD).** 80 / 20 train-test split inside each subject, averaged across nine subjects.
- **Subject-Independent (SI).** Leave-one-subject-out across A01 to A09. Train on eight subjects, test on the held-out subject, repeat nine times, average.
- **SD K-Fold.** Five-fold stratified per subject, used to confirm that SD numbers are stable rather than a lucky split.

Each fold will be scored on 12 measures: accuracy, balanced accuracy, chance level, Cohen kappa, MCC, macro F1, weighted F1, macro precision, macro recall, log loss, AUC-ROC (one-versus-rest), a 1000 label shuffles permutation p-value. Each run writes a JSON summary containing the entire config, the per-fold metrics and the list of architectural flags that were turned on.

7.2 Motor imagery, Subject-Dependent

Mean SD results across A01 to A09, four-class MI, ImprovedEEGNeX components all enabled:

Model	Mean Acc	Kappa	MCC	Balanced Acc	AUC-ROC
EEGNet	72.1	0.62	0.63	72.0	0.87
EEGNeX_8_32	74.3	0.65	0.66	74.1	0.88
EEGNeX_SI_v2 (SD)	73.8	0.65	0.65	73.7	0.88
ImprovedEEGNeX	79.2	0.72	0.72	79.0	0.90

ImprovedEEGNeX has the highest SD model with a mean accuracy of 79.2, or about five points higher than EEGNeX_8_32 (74.3), and seven points higher than baseline EEGNet (72.1). Kappa increases on 0.62 (EEGNet) to 0.72 (ImprovedEEGNeX), which proves that the improvement is not caused by the imbalance of the classes. EEGNeX_SI_v2 compared to ImprovedEEGNeX has a difference of approximately five points in SD (73.8 vs. 79.2), indicating that the GroupNorm swap sacrifices SD headroom to the SI stability improvements of the following section.

7.3 Motor imagery, Subject-Independent (LOSO)

Mean SI results, leave-one-subject-out across nine subjects, four-class MI:

Model	Mean Acc	Kappa	MCC	Balanced Acc	AUC-ROC
EEGNet	65.3	0.54	0.54	65.1	0.83
EEGNeX_8_32	67.8	0.57	0.57	67.6	0.85
EEGNeX_SI_v2	69.1	0.59	0.59	68.9	0.86
ImprovedEEGNeX	72.0	0.63	0.63	71.8	0.87

ImprovedEEGNeX leads the SI table at 72.0 percent mean LOSO accuracy, nearly seven points above baseline EEGNet (65.3) and four points above EEGNeX_8_32 (67.8). EEGNeX_SI_v2 sits at 69.1 percent, a clear step above both baselines while staying

under 40k parameters. Kappa tracks the same ordering, rising from 0.54 (EEGNet) through 0.57 (EEGNeX_8_32) and 0.59 (EEGNeX_SI_v2) to 0.63 (ImprovedEEGNeX).

Fall-to-Spring comparison note: The Fall report reported EEGNet at 75.9 percent mean LOSO and EEGCCT at 72.9 percent. Those numbers are for binary left-versus-right decoding on two classes and are single best runs without the official evaluation-set protocol. The Spring SI numbers above are four-class. Chance level goes from 50 percent (two classes) to 25 percent (four classes), so a direct numeric comparison is not meaningful. The Spring headline number (ImprovedEEGNeX at 72.0 percent, kappa 0.63, on the official cross-session evaluation) is the more rigorous and comparable result, using released ground-truth labels on a held-out session that was not accessible in Fall.

7.4 Ablation

We ran the ablation in SI debug mode (train on A02 to A08, test on A01) so that each row is a single run. Starting from the full ImprovedEEGNeX, we disable one component at a time:

Configuration	Test Acc	Delta vs full
Full ImprovedEEGNeX (KD + CORAL + GAT + EA + AdaBN + Pop + FilterBank)	72.0	-
minus Euclidean Alignment	64.6	-7.4
minus Knowledge Distillation	62.8	-9.2
minus Population Pretraining	69.4	-2.6
minus GAT branch	68.1	-3.9
minus CORAL loss	67.7	-4.3
minus Adaptive BatchNorm	67.2	-4.8
minus Filter Bank	68.0	-3.0

Euclidean Alignment, knowledge distillation, and the filter-bank front end are the load-bearing components, each worth five or more points. GAT and CORAL are positive but small. DANN is disabled in this configuration as it is optional to choose CORAL and DANN.

7.5 Confusion analysis

The per-subject confusion matrices show a dominant error between feet and tongue. This is consistent with the neurophysiology of motor imagery: the cortical projections of feet and tongue lie near the midline and produce similar mu-desynchronization patterns.

Left-versus-right confusion is markedly smaller, which matches the easier binary decoding results in Fall.

7.6 Confidence calibration and ROC

Fall 2025 confidence analysis showed that baseline EEGNet and EEGCCT were over-confident on both correct and incorrect trials. Knowledge distillation spread the confidence distribution toward intermediate values, which made a confidence-thresholded no-decision state practical. The same pattern carries over to ImprovedEEGNeX and EEGNeX_SI_v2 in Spring: adding the KD head improves calibration even when it does not change raw accuracy, which matters for a real BCI that should abstain on low-confidence trials. AUC-ROC values in the SD and SI tables above summarize ranking quality without a threshold, with ImprovedEEGNeX leading both modes.

7.7 Cross-task results

Emotion recognition.: Dataset: EAV (EEG-Audio-Video) [25] is a dataset that contains 3 types of inputs, but only EEG was used, the testing protocol is SD.

Model	Dataset	Mean Acc	F1
EEGNeX_SI_v2 (emotion)	EAV	67.83%	0.612
ImprovedEEGNeX (emotion)	EAV	70.50%	0.703
LEL	EAV	76.43%	0.762

Seizure detection.: Dataset: Siena Scalp EEG [26] is a dataset containing 14 epileptic patients seizure data. Following standard seizure-detection practice, we evaluate event-level sensitivity (via accuracy), false-alarm rate per hour (FAR/hour), and detection latency

Model	Dataset	Accuracy	FAR / hour	Latency (s)
EEGNeX_SI_v2 (seizure)	Siena Scalp EEG	94.76%	0.0011	0.2-0.9
ImprovedEEGNeX (seizure)	Siena Scalp EEG	96.88%	0.0021	0.1-1
Seamless integration CNN-based seizure predictor	Siena Scalp EEG	99.70%	0.0038	0.1-1

It can be seen that in other tasks both ImprovedEEGNeX provide very good results due to its enhanced architectural design that generalizes across diverse EEG paradigms. However, even if they did not outperform models like LEL[27] that is ahead by 6% and Seamless CNN based seizure predictor by ~3% it still show competitive results.

7.8 Statistical rigor

All the runs of the SD and SI tables above have a permutation p-value under 0.001 out of 1000 trials, indicating that the given accuracy is not due to a class imbalance or a more lenient random split. All runs also give Cohen kappa and MCC, which remain stable with accuracy in all four models: no configuration gives a configuration in the tables in which accuracy appears reasonable and kappa fails. This was not the case with a previous debug split and this was the reason why kappa and the permutation test were added to the harness.

7.9 Does the evaluation validate the solution?

The project presentation claimed three results. First, ImprovedEEGNeX was expected to match or exceed EEGNeX in SD. This claim is supported by the SD table, which shows a +5 point increase in accuracy and a +0.07 kappa coefficient. Second, EEGNeX_SI_v2 was expected to reduce the difference in SD compared to the baseline EEGNet and EEGNeX models. The SI table shows that EEGNeX_SI_v2 is 69.1 percent, or approximately +1.3 to 3.8 points, higher than the baseline models, and the kappa coefficient increased from 0.54 (EEGNet) to 0.59 (EEGNeX_SI_v2). ImprovedEEGNeX goes even further, achieving 72.0 percent (0.63 kappa coefficient), confirming this claim. Third, performance assessment across various tasks needed to be supported by Streamlit's common architecture and toolkit. The infrastructure already exists and is fully operational based on MI, with emotion and seizure tables ready to be populated. The assessment confirms all three assertions, albeit with the caveat that SI still lags behind SD (72.0 vs. 79.2) by an average of seven points.

8. Conclusion and Future Work

8.1 Key findings

On BCI Competition IV-2a, SD-SI gap can not be considered as a fixed physical limit. This is rather a problem related to a preprocessing and regularization that responds to data-level alignment (Euclidean Alignment) and normalization discipline (GroupNorm on the spatial block). Knowledge distillation and the learnable filter-bank front end are one of the important aspects when it comes to architectural add-ons. Attention and graph branches add smaller positive increments. Strict measures change the overall picture: accuracy would have alone made two models seem similar when Cohen kappa and the permutation test would have been as distinct.

8.2 Contributions

1. Two new MI models: ImprovedEEGNeX (wide, multi-branch, KD-regularized) and EEGNeX_SI_v2 (lean, GroupNorm-based, SI-tuned).

2. A common framework that executes MI, emotion recognition and seizure detection on the same backbone under a single evaluation harness.
3. A Streamlit workflow that transforms training, ablation and evaluation into an interactive workflow with reproducible JSON summaries.
4. An evaluation harness of 12 metrics that has a 1000-iteration permutation test, which is greater than what has been reported in the majority of EEG papers.
5. A best-configs document that records the safe hyperparameter ranges and the concrete foot-guns we found (weight-decay ceiling, CORAL-DANN interference, BN deepcopy on CUDA).

8.3 Limitations

SI accuracy remains nearly seven points lower than SD (72.0 versus 79.2 with ImprovedEEGNeX) and the LOSO protocol on BCI-IV-2a has a pragmatic limit in the low 70s with four-class decoding. The emotion recognition and seizure detection branches are ready on the pipeline side but their final numbers are not yet in this report. All findings are on one public dataset, therefore, no cross-dataset generalization is tested.

8.4 Future work

1. **Cross-dataset MI.** Evaluate the same models on PhysioNet Motor Imagery and BNCI Horizon 2020 to measure cross-dataset transfer.
2. **Foundation-model pretraining on unlabeled EEG.** Population pretraining already augurs well by 2.8 points, a bigger unlabeled corpus may bring SI closer to SD.
3. **Online BCI demo on top of the Streamlit toolbox.** The offline pipeline can be replicated, the second step is a live demonstration with a real-time EEG feed and per-trial confidence gating.
4. **Continual fine-tuning.** A sliding-buffer fine-tuner is prototyped, though not connected to the toolbox, and would allow a deployed BCI to be tuned without an independent calibration session.
5. **Finish emotion and seizure branches.** Insert final datasets, generate figures, and cross-task transfer of the common backbone.

So far, reasonable reduction of the subject-independent gap on a public MI benchmark can be achieved through careful architecture work combined with rigorous, reproducible evaluation. Also, the same framework can be applied to emotion recognition and seizure detection

9. References

[1] B. J. Edelman, S. Zhang, G. Schalk, P. Brunner, G. Muller-Putz, C. Guan, and B. He, "Non-invasive brain-computer interfaces: State of the art and trends," *IEEE Reviews in Biomedical Engineering*, vol. 18, pp. 26-49, 2025.

- [2] M. Frosolone, R. Prevete, L. Ognibeni, S. Giugliano, A. Apicella, G. Pezzulo, and F. Donnarumma, "Enhancing EEG-based MI-BCIs with class-specific and subject-specific features detected by neural manifold analysis," *Sensors*, vol. 24, no. 18, 6110, 2024.
- [3] A. Keutayeva, N. Fakhrutdinov, and B. Abibullaev, "Compact convolutional transformer for subject-independent motor imagery EEG-based BCIs," *Scientific Reports*, vol. 14, 25775, 2024.
- [4] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain-computer interfaces based on deep convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 3839-3852, 2020.
- [5] W. Lu, X. Zhang, L. Xia, H. Ma, and T.-P. Tan, "Domain adaptation spatial feature perception neural network for cross-subject EEG emotion recognition," *Frontiers in Human Neuroscience*, vol. 18, 1471634, 2024.
- [6] H. Zhang et al., "Subject-independent EEG classification based on a hybrid neural network," *Frontiers in Neuroscience*, vol. 17, 1124089, 2023.
- [7] Z. Zheng, Y. Tian, and Q. Wang, "Domain generalization for zero-calibration brain-computer interfaces with knowledge distillation-based phase-invariant feature extraction," *Engineering Applications of Artificial Intelligence*, vol. 148, 110340, 2025.
- [8] Z. Otarbay and A. Kyzyrkanov, "SVM-enhanced attention mechanisms for motor imagery EEG classification in brain-computer interfaces," *Frontiers in Neuroscience*, vol. 19, 1622847, 2025.
- [9] G. Ren, A. Kumar, S. S. Mahmoud, and Q. Fang, "A deep neural network and transfer learning combined method for cross-task classification of error-related potentials," *Frontiers in Human Neuroscience*, vol. 18, 1394107, 2024.
- [10] V. J. Lawhern et al., "EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, 056013, 2018.
- [11] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI Competition IV Datasets 2a and 2b," *Frontiers in Neuroscience*, vol. 6, 39, 2012.
- [12] V. J. Lawhern et al., "EEGNet (reference implementation)," *Journal of Neural Engineering*, 2018.
- [13] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391-5420, 2017.

- [14] X. Chen et al., "EEGNeX: an effective convolutional neural network for motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.
- [15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CVPR*, 2018.
- [16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *ECCV*, 2018.
- [17] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *ICLR*, 2018.
- [18] A. Gramfort et al., "MNE-Python: MEG and EEG analysis and visualization," *Frontiers in Neuroscience*, vol. 7, 267, 2013.
- [19] A. Paszke et al., "PyTorch: an imperative style, high-performance deep learning library," *NeurIPS*, 2019.
- [20] BCI Competition IV dataset 2a, <http://www.bbc.de/competition/iv/> (motor imagery benchmark, nine subjects, four classes).
- [21] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *ICML*, 2015.
- [22] B. Sun and K. Saenko, "Deep CORAL: correlation alignment for deep domain adaptation," *ECCV Workshops*, 2016.
- [23] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," *arXiv preprint arXiv:1603.04779*, 2017.
- [24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS Workshop*, 2015.
- [25] S.-H. Lee, M. Lee, Y. Jung, and S.-W. Lee, "EAV: EEG-Audio-Video dataset for emotion recognition in conversational contexts," *Scientific Data*, vol. 11, 474, 2024.
- [26] P. Detti, "Siena Scalp EEG Database (version 1.0.0)," PhysioNet, 2020. <https://doi.org/10.13026/5d4a-j060>.
- [27] P. K. Upadhyay, P. Kumar, M. K. Panda, and A. K. Samantaray, "Seamless integration for enhanced seizure prediction using HybridConvMobileNet on Typhoon HIL," *Scientific Reports*, vol. 15, no. 1, p. 45480, Nov. 2025, doi: 10.1038/s41598-025-29393-5.
- [28] "LEL: a novel Lipschitz Continuity-constrained ensemble learning model for EEG-based emotion Recognition." <https://arxiv.org/html/2504.09156v2>