Metaheuristic Algorithms with Applications

www.maa.reapress.com

Metaheur. Algor. Appl. Vol. 1, No. 3 (2024) 269–292.

Paper Type: Original Article

Metaheuristic-Guided Neural Architecture Search for Cross-Session EEG Speech Imagery Decoding

Mahmoud Rahimi[1,*], Sara Al-Farsi[2]

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran; m.rahimi@sharif.edu.

[2] Department of Computer Science and Software Engineering, College of Information Technology, UAE University, Al Ain, United Arab Emirates; s.alfarsi@uaeu.ac.ae.

Citation:

Received: 19 January 2024 *Revised: 28 June 2024* *Accepted: 02 July 2024*	Rahimi, M., & Al-Farsi, S. (2024). Metaheuristic-guided neural architecture search for cross-session EEG Speech Imagery decoding. *Metaheuristic algorithms with applications, 1*(3), 269-292.

Abstract

Brain-Computer Interfaces (BCIs) based on Speech Imagery (SI) hold transformative potential for restoring communication in individuals with locked-in syndrome, Amyotrophic Lateral Sclerosis (ALS), and severe dysarthria. However, practical deployment of Speech Imagery BCIs is critically limited by the cross-session transfer problem: Electroencephalography (EEG) signal distributions shift substantially between recording sessions due to electrode repositioning, impedance fluctuations, and cognitive state variations, causing within-session classification accuracies of 60–70% to collapse to 30–40% in cross-session evaluation. Existing deep learning architectures for EEG decoding are hand-designed and optimized for within-session performance, leaving the cross-session generalization problem largely unaddressed. We propose Metaheuristic-guided Neural Architecture Search (NAS) for Speech Imagery (MetaNAS-SI), a novel framework that automatically discovers optimal neural network architectures explicitly optimized for cross-session EEG Speech Imagery decoding. MetaNAS-SI introduces a hybrid evolutionary search combining aging evolution for macro-architecture exploration with Differential Evolution (DE) for micro-architecture optimization, a cross-session fitness function based on Leave-One-Session-Out (LOSO) evaluation, a searchable Session-Invariant Feature Alignment Module (SIFAM) that performs domain adaptation via maximum mean discrepancy within candidate architectures, and an EEG-specific search space incorporating physiologically motivated temporal convolutions, Common Spatial Pattern (CSP)-inspired spatial filters, and multi-scale temporal attention. MetaNAS-SI was evaluated on three Speech Imagery datasets: KaraOne (5-class), the coretto imagined speech dataset (4-class), and an in-house Farsi-Arabic dataset (6-class, 5 sessions per subject). Cross-session accuracy reached 45.8% on KaraOne (vs. 36.4% for the best baseline), 52.3% on Coretto (vs. 43.8%), and 48.1% on the Farsi-Arabic dataset (vs. 35.2%). Discovered architectures were 3–15× smaller than hand-designed alternatives while achieving 18–38% relative improvement in cross-session accuracy. Ablation analysis confirmed the critical contributions of the cross-session fitness function and SIFAM module. Pareto-optimal architecture selection enables deployment-aware trade-offs between accuracy, model size, and inference latency for real-time BCI applications.

Keywords: Brain-computer interface, Speech imagery, Electroencephalography decoding, Neural architecture search, Metaheuristic optimization, Cross-session transfer, Domain adaptation, Evolutionary optimization.

Corresponding Author: m.rahimi@sharif.edu

https://doi.org/10.48313/maa.v1i3.101

1 | Introduction

The aspiration to decode human thought directly from neural activity has driven decades of research in Brain-Computer Interface (BCI) technology. Among the most compelling applications is the development of silent speech interface systems capable of translating imagined speech (Speech Imagery) into text or synthesized audio without any overt vocalization or muscular movement. Such interfaces would be transformative for individuals who have lost the ability to speak due to neurological conditions including locked-in syndrome, Amyotrophic Lateral Sclerosis (ALS), brainstem stroke, and severe dysarthria [1], [2]. For these individuals, conventional assistive communication devices that rely on residual motor function are often inadequate, making neural decoding of speech intention a critically needed alternative.

Electroencephalography (EEG) has emerged as the preferred non-invasive modality for BCI applications due to its portability, relatively low cost, high temporal resolution (millisecond scale), and lack of surgical risk [3]. However, EEG signals are inherently noisy, suffer from low spatial resolution due to volume conduction through the skull, and exhibit substantial variability across recording sessions, subjects, and even trials within a single session [4]. These characteristics pose fundamental challenges for reliable neural decoding, particularly for Speech Imagery, which produces weaker and more spatially diffuse cortical activation patterns compared to other BCI paradigms such as motor imagery.

Speech Imagery engages a distributed cortical network spanning Broca's area (inferior frontal gyrus), Wernicke's area (posterior superior temporal gyrus), the supplementary motor area, primary and premotor cortex, and auditory association cortex [5], [6]. Unlike motor imagery, which produces relatively focal and stereotyped sensorimotor rhythms (mu and beta band modulations over the motor strip), Speech Imagery activates bilateral and diffuse cortical regions with subtle oscillatory signatures that are difficult to distinguish from background EEG noise. Consequently, classification accuracy for multi-class imagined speech typically ranges from 30% to 50% even in within-session evaluations, substantially lower than the 70–90% accuracy commonly achieved in motor imagery Brain-Computer Interfaces (BCIs) [7–9].

The most critical barrier to real-world deployment of Speech Imagery BCIs is the cross-session transfer problem. EEG signal distributions shift systematically between recording sessions due to multiple sources of non-stationarity: Electrode repositioning introduces spatial inconsistencies, electrode-scalp impedance varies with skin preparation and environmental conditions, the subject's cognitive state, alertness, and mental strategy may change across sessions, and neural adaptation or learning effects alter the underlying neural responses over time [3], [10], [11]. These distributional shifts cause within-session classification accuracies of 60–70% to degrade to 30–40% when a classifier trained on one session is applied to data from a different session [12]. This dramatic performance collapse renders current Speech Imagery BCI systems impractical for daily use, as recalibration at each session is prohibitively burdensome.

Deep learning approaches have shown promise in EEG-based BCI decoding. EEGNet [13], a compact convolutional neural network employing depthwise and separable convolutions, demonstrated cross-paradigm generalization capabilities. DeepConvNet and ShallowConvNet [14] explored deeper and shallower architectural variants for EEG motor imagery classification. More recently, EEG-conformer [15] combined convolutional and transformer components for enhanced temporal feature extraction. However, all these architectures share a common limitation: they are hand-designed by domain experts through laborious trial-and-error experimentation, and none were explicitly optimized for cross-session generalization in Speech Imagery decoding. Their architectural choices, filter sizes, network depth, and attention mechanisms were selected based on within-session performance on motor imagery or event-related potential tasks, with no guarantee of transferability to the distinct characteristics of Speech Imagery signals or cross-session robustness.

Neural Architecture Search (NAS) offers a principled approach to automating the discovery of optimal network architectures for a given task. Seminal NAS methods include reinforcement learning-based search [16], differentiable architecture search [17], and evolutionary approaches such as aging evolution [18]. NAS

has achieved remarkable success in computer vision and natural language processing, often discovering architectures that surpass human-designed models. However, the application of NAS to EEG-based BCI remains in its infancy, with only a handful of studies exploring NAS for motor imagery classification [19], [20]. Critically, no existing NAS framework has been designed for Speech Imagery EEG decoding, and none have incorporated cross-session generalization as an explicit optimization objective.

To address this significant research gap, we propose Metaheuristic-guided NAS for Speech Imagery (MetaNAS-SI), a comprehensive framework that automatically discovers optimal deep neural network architectures for cross-session EEG Speech Imagery decoding. MetaNAS-SI introduces five key contributions:

I. Hybrid evolutionary NAS combining aging evolution with Differential Evolution (DE): a two-level search strategy that uses aging evolution for macro-architecture exploration (layer types, skip connections, network topology) and DE for micro-architecture optimization (filter sizes, dropout rates, learning rates), efficiently navigating the vast combinatorial architecture space.

II. Cross-session fitness function: unlike standard NAS that optimizes single-session accuracy, MetaNAS-SI evaluates candidate architectures using a Leave-One-Session-Out (LOSO) protocol, explicitly selecting architectures that generalize across sessions rather than overfitting to session-specific artifacts.

III. Session-Invariant Feature Alignment Module (SIFAM): a searchable architectural module embedded within candidate networks that performs domain adaptation between sessions using Maximum Mean Discrepancy (MMD) loss, with the NAS optimizing SIFAM's internal structure alongside the main classifier architecture.

IV. EEG-specific search space: a neurophysiologically informed search space incorporating temporal convolutions with filter lengths matched to canonical EEG frequency bands, Common Spatial Pattern (CSP)-inspired learnable spatial filters, and multi-scale temporal attention mechanisms.

V. Pareto-optimal architecture selection: multi-objective optimization balancing cross-session accuracy, model parameter count, and inference latency, returning a Pareto front of architectures suitable for different deployment scenarios.

We evaluate MetaNAS-SI on three Speech Imagery EEG datasets: The KaraOne dataset [21], the Coretto Imagined Speech dataset [22], and an in-house Farsi-Arabic Imagined Speech (FA-IS) dataset collected from 12 subjects across 5 sessions each. Our results demonstrate that MetaNAS-SI achieves cross-session accuracies of 45.8%, 52.3%, and 48.1% on these datasets, respectively, substantially outperforming 10 baseline methods while producing architectures that are 3–15× more compact than hand-designed alternatives.

2 | Related Work

2.1 | Speech Imagery Electroencephalography Decoding

The decoding of imagined speech from EEG has been investigated using both traditional machine learning and deep learning approaches. Early work by DaSalla et al. [7] demonstrated the feasibility of classifying two imagined vowels (/a/ vs. /u/) using EEG with CSP and Support Vector Machines (SVM), achieving approximately 70% binary classification accuracy within a single session. Brigham and Kumar [8] extended this to imagined syllables using spectral features and achieved above-chance classification for binary tasks. Nguyen et al. [9] employed Filter Bank Common Spatial Pattern (FBCSP) combined with regularized linear discriminant analysis for multi-class imagined speech, reporting accuracies in the range of 35–50% for 4–5 classes within-session.

Power Spectral Density (PSD) features in theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz) frequency bands have been identified as informative for imagined speech discrimination [23]. Wavelet-based features have also shown utility, capturing transient spectral dynamics associated with the temporal unfolding of imagined speech [24]. However, these feature-engineering approaches require domain expertise, may not

capture complex nonlinear interactions in the EEG data, and have shown limited generalization to new sessions or subjects.

Deep learning methods have more recently been applied to Speech Imagery EEG. Saha et al. [24] applied convolutional neural networks to classify imagined speech and demonstrated improvements over traditional approaches but noted substantial performance degradation in cross-subject evaluation. Lee et al. [25] explored Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks for temporal modeling of imagined speech EEG, showing that sequential modeling could capture temporal dynamics but at the cost of computational complexity. EEGNet [13] has been widely adopted as a baseline architecture for various EEG classification tasks, including Speech Imagery, owing to its compact design and cross-paradigm versatility. Attention mechanisms have been explored by Nieto et al. [26], who demonstrated that temporal attention could selectively weight informative time segments in imagined speech trials, improving classification of imagined words. Despite these advances, the state-of-the-art for multi-class (4–6 classes) imagined speech classification remains at 30–50% within-session accuracy and 25–40% cross-session, highlighting the fundamental difficulty of this decoding task.

2.2 | Cross-Session and Cross-Subject Electroencephalography Transfer

The problem of distributional shift between EEG recording sessions has been extensively studied in the motor imagery BCI literature, though far less so for Speech Imagery. Wu et al. [27] provided a comprehensive survey of transfer learning methods for EEG-based BCIs, categorizing approaches into instance-based, feature-based, and model-based transfer. He and Wu [28] proposed Euclidean Alignment (EA), a simple but effective method that aligns the mean covariance matrix of EEG trials across sessions or subjects to a common reference, demonstrating substantial improvements in cross-session motor imagery classification. Barachant et al. [29] developed Riemannian geometry-based approaches that operate on the manifold of symmetric positive definite matrices (EEG covariance matrices), providing geometrically principled alignment and classification that is inherently robust to affine transformations of the EEG data.

Deep domain adaptation methods have also been applied to BCI transfer. Ozdenizci et al. [30] employed adversarial domain adaptation for cross-subject EEG transfer, training a domain discriminator alongside the main classifier to learn domain-invariant features. Long et al. [31] introduced deep adaptation networks using MMD for domain adaptation in general machine learning, which has subsequently been adopted for EEG applications. Zhao et al. [32] proposed multi-source domain adaptation for motor imagery BCI, aggregating information from multiple source subjects to improve transfer to a new user. Session-to-session style transfer and normalization techniques [33], [34] have shown promise in reducing cross-session variability, but these methods have been predominantly validated on motor imagery tasks with relatively strong and focal neural signals. The application of domain adaptation to Speech Imagery EEG, where the signals are weaker and more distributed, remains largely unexplored.

2.3 | Neural Architecture Search

NAS automates the process of designing neural network architectures, replacing the conventional approach of manual architecture engineering. The field was catalyzed by Barret and Le Quoc [16], who used reinforcement learning to train a controller RNN to generate neural architecture descriptions, achieving competitive performance on image classification benchmarks albeit at enormous computational cost (800 GPU-days). Subsequent work dramatically reduced this cost: DARTS [17] reformulated architecture search as a continuous optimization problem by relaxing the discrete architecture space, enabling gradient-based search that completes in a single GPU-day. Real et al. [18] demonstrated that simple evolutionary methods, particularly aging evolution, which removes the oldest individuals from the population regardless of fitness, could match or exceed the performance of reinforcement learning-based NAS with better computational efficiency and simplicity.

Hardware-aware NAS methods [27], [35] introduced multi-objective optimization to jointly optimize accuracy and hardware efficiency metrics (latency, FLOPs, parameter count), producing models suitable for deployment on resource-constrained devices. One-shot NAS methods [36], [37] amortize the cost of architecture evaluation by training a single supernet that subsumes all candidate architectures, enabling rapid architecture scoring without individual training. These advances in general NAS have not been systematically translated to the EEG/BCI domain.

NAS for EEG remains nascent. Sun et al. [19] proposed NASIL for motor imagery EEG classification, applying a simplified NAS procedure with a limited search space, but did not address cross-session transfer or Speech Imagery. Zhu et al. [20] introduced AutoEEG, an automated model design framework for general EEG classification that employed differentiable search, but focused on within-subject evaluation and did not incorporate domain adaptation or cross-session objectives. To our knowledge, no NAS framework has been proposed for Speech Imagery EEG decoding, and no NAS approach in the BCI domain has incorporated cross-session generalization as an explicit optimization target or included searchable domain adaptation modules within the architecture space.

2.4 | Metaheuristic Optimization for Neural Networks

Metaheuristic optimization encompasses a family of nature-inspired, population-based algorithms that search for optimal solutions in complex, high-dimensional spaces without requiring gradient information. Key algorithms include Genetic Algorithms (GAs) [38], particle swarm optimization [39], DE [40], simulated annealing [41], and ant colony optimization [42]. These methods suit the combinatorial nature of the architecture search problem, where the search space is discrete, non-differentiable, and multi-modal.

Neuroevolution, the application of evolutionary algorithms to optimize neural network architectures, weights, or both, has a rich history [43]. NeuroEvolution of Augmenting Topologies (NEAT) [44] evolved both topology and weights of neural networks through complexification. More recently, Large-Scale Evolution of image classifiers [45] and AmoebaNet [18] demonstrated that evolutionary methods can discover competitive architectures at scale. In the BCI domain, metaheuristic optimization has been applied primarily for feature selection and hyperparameter tuning: GAs for CSP filter selection [46], particle swarm optimization for SVM hyperparameters [47], and DE for optimizing neural network weights in EEG classification [48]. However, full neural architecture optimization via metaheuristics for EEG decoding remains unexplored, particularly with EEG-specific architectural components and cross-session objectives. MetaNAS-SI bridges this gap by casting the cross-session Speech Imagery architecture design problem as a multi-objective metaheuristic optimization problem with an EEG-tailored search space.

3 | Methodology

3.1 | Problem Formulation

Let a Speech Imagery EEG dataset consist of recordings from S sessions for a given subject, where each session s contains N_s labeled trials of K-class imagined speech. Each trial is represented as a matrix $X \in \mathbb{R}^{C\times T}$, where C denotes the number of EEG channels, and T denotes the number of temporal samples. The associated class label is $y \in \{1, 2, \ldots, K\}$.

The objective of MetaNAS-SI is to discover an architecture A* and corresponding trained weights W* that maximize cross-session classification performance. We formalize this through a LOSO protocol:

$$A^* = \arg\max_{A \in \mathcal{A}} (1/S) \Sigma_{s=1}^{S} \mathrm{Acc}(A, W_A, D_{test}^{s} \mid D_{train}^{-s}), \quad (1)$$

where $D_{train}^{-s} = \{D^1, \ldots, D^{s-1}, D^{s+1}, \ldots, D^S\}$ denotes the training data comprising all sessions except session s, and D_{test}^{s} denotes the test data from the held-out session s. This LOSO cross-session accuracy serves as the primary fitness function for the NAS procedure, directly rewarding architectures that generalize across sessions.

Simultaneously, we seek to minimize the model parameter count $|A|$ (for efficient real-time BCI deployment) and inference latency $L(A)$. The complete multi-objective formulation is:

$$\text{maximize } \{\text{Acc}_{\text{LOSO}}(A), -|A|, -L(A)\}. \quad (2)$$

Eq. (2) yields a Pareto front of non-dominated architectures, from which deployments can select based on application-specific constraints (Section 3.6).

3.2|Electroencephalography-Specific Search Space

A central contribution of MetaNAS-SI is the design of a search space that incorporates neurophysiologically informed operations tailored to the characteristics of EEG signals. The search space consists of four categories of operations, a macro-architectural skeleton, and the SIFAM module.

3.2.1|Temporal convolution operations

Temporal convolutions are the primary feature extractors in EEG deep learning, and their filter lengths directly determine the frequency bands they are sensitive to. We define temporal convolution operations whose kernel sizes are explicitly matched to the canonical EEG frequency bands at a sampling rate of 256 Hz:

Table 1. Temporal convolution kernels matched to canonical EEG frequency bands.

Operation	Kernel Size (Samples)	Temporal Extent (ms)	Target Frequency Band
TC-Delta	64	250	Delta (0.5–4 Hz)
TC-Theta	32	125	Theta (4–8 Hz)
TC-Alpha	26	100	Alpha (8–13 Hz)
TC-Beta	13	50	Beta (13–30 Hz)
TC-Gamma	7	25	Gamma (30–100 Hz)
TC-multi	7, 13, 26, 32, 64 (parallel)	Multi-scale	All bands (concatenated)

The physiological rationale is that different aspects of Speech Imagery are encoded in different frequency bands: Beta desynchronization in motor planning areas [49], gamma activation in language processing regions [50], theta oscillations in working memory and semantic processing [51], and alpha suppression in sensory and motor cortex [52]. The TC-Multi operation applies all kernel sizes in parallel and concatenates their outputs, enabling multi-scale temporal feature extraction within a single layer.

3.2.2|Spatial operations

Spatial operations capture the spatial distribution of neural activity across EEG channels. We define four spatial operation types:

I. SC-full: depthwise separable spatial convolution across all channels, following the EEGNet paradigm [13]. Such a configuration learns unconstrained spatial filters that linearly combine channel activities.

II. SC-CSP: a learnable CSP-inspired spatial filter, parameterized as $Z = W^T X$, where $W \in \mathbb{R}^{C \times C'}$ is a learnable spatial projection matrix and $C' < C$ is the number of output spatial filters (searchable: 4, 8, 16, or 32). Such an embedding embeds the CSP principle of variance-based spatial filtering within the end-to-end trainable network.

III. SC-region: region-based spatial convolution that applies separate spatial filters to anatomically defined channel groups: frontal (Fp, F channels), temporal (T channels), central-parietal (C, P channels), and occipital (O channels) before concatenation. Such a design enforces anatomical prior knowledge about the regional organization of speech-related cortical activity.

IV. SC-attention: a channel attention mechanism following the squeeze-and-excitation paradigm [53], which learns to dynamically weight EEG channels based on their task-relevance within each trial.

3.2.3 | Temporal 1

Temporal attention mechanisms enable the network to selectively attend to informative time segments within an EEG trial:

I. TA-self: multi-head self-attention over the temporal dimension with the number of attention heads searchable among {1, 2, 4, 8}. Such an approach captures long-range temporal dependencies in the EEG signal.

II. TA-causal: causal (masked) temporal attention that restricts attention to past and present time points, preserving the temporal ordering of the Speech Imagery process.

III. TA-relative: relative position-encoded attention [54] that incorporates positional information to encode the temporal structure of neural events.

3.2.4 | Auxiliary operations

The search space also includes standard neural network operations: identity (skip connections), average pooling with searchable kernel sizes ({2, 4, 8}), batch normalization followed by ELU activation, dropout with searchable rates ({0.1, 0.25, 0.5}), and a zero operation (representing no connection). These provide the NAS with flexibility to construct architectures of varying depth and connectivity.

3.2.5 | Macro-architecture

The macro-architecture consists of up to 8 cells (layers), where each cell selects one operation from the pool defined above. The output channel dimension for each cell is searchable among {16, 32, 64, 128}. Binary skip connections may be established between any pair of cells (represented as a binary adjacency matrix). The final classification head consists of global average pooling followed by a dense layer with softmax activation producing K-class probabilities.

3.3 | Hybrid Evolutionary Neural Architecture Search Algorithm

MetaNAS-SI employs a two-level hybrid evolutionary search strategy that decomposes the architecture optimization into macro-architecture search (discrete, combinatorial) and micro-architecture optimization (continuous, numerical), each addressed by an appropriate metaheuristic algorithm.

3.3.1 | Level 1: macro-architecture search via aging evolution

The macro-architecture defines the computational graph of the network: The sequence of operations in each cell, skip connections between cells, SIFAM placement, and output channel dimensions. The resulting formulation is a combinatorial optimization problem over a discrete space, well-suited to evolutionary search. We adopt the aging evolution strategy [18], which maintains population diversity through an age-based removal mechanism rather than fitness-based selection pressure alone.

The algorithm maintains a population of P=25 architecture candidates, each represented as a genome encoding: 1) the operation type for each of the up to 8 cells, 2) the output channel dimension for each cell, 3) the binary skip connection adjacency matrix, and 4) the SIFAM position (after cell 2, 4, or 6). In each generation, S=5 individuals are randomly sampled from the population (tournament selection); the fittest among them is selected as the parent, and a child is produced by applying a single mutation. The child is added to the population, and the oldest individual is removed regardless of fitness. This aging mechanism prevents premature convergence by continuously injecting diversity and preventing long-lived dominant genotypes from monopolizing the population.

Mutation operations include: 1) changing one cell's operation type (uniformly sampled from the operation pool), 2) adding or removing a skip connection between two randomly selected cells, 3) changing the SIFAM position to a different cell, and 4) changing one cell's output channel dimension. Each mutation is equally likely, and exactly one mutation is applied per child.

Fitness evaluation at *Level 1* uses a computationally efficient proxy: Each candidate architecture is trained for 20 epochs on a reduced dataset comprising only 2 sessions with a LOSO split (train on 1 session, test on the other, then swap and average). This proxy training takes approximately 15 minutes per architecture on an NVIDIA RTX 3090 GPU. The search runs for 200 generations, evaluating approximately 200 unique architectures. Total *Level 1* cost: approximately 50 GPU-hours.

3.3.2 | Level 2: micro-architecture optimization via differential evolution

For the top K = 5 macro-architectures identified in *Level 1*, we optimize continuous hyperparameters using DE [40]. The continuous parameter vector θ includes: Dropout rates for each cell, learning rate, weight decay, SIFAM alignment strength λ_{MMD}, batch size, and data augmentation strength. DE is well-suited to this continuous, bounded optimization problem due to its simplicity, few control parameters, and robust global search capability.

We employ the DE/rand/1/bin variant with a population of $N_{DE} = 20$ individuals, scale factor F = 0.5, and crossover rate CR = 0.9. Each DE generation produces a trial vector through mutation and binomial crossover, which replaces the target individual if it achieves higher fitness. Fitness at *Level 2* is evaluated using the full LOSO cross-session protocol across all available sessions with complete training (100 epochs with early stopping). The DE runs for 50 generations per macro-architecture. With early stopping and fitness caching (identical or very similar parameter vectors are not re-evaluated), the total *Level 2* cost is approximately 40 GPU-hours.

The combined two-level search cost is approximately 90 GPU-hours. While substantial, this is a one-time investment per dataset and BCI paradigm; the discovered architecture can subsequently be trained for new subjects in less than 2 hours.

Algorithm 1. MetaNAS-SI two-level architecture search.

Input: EEG dataset D = {D^1, ..., D^S} (S sessions), search space A, population sizes P=25 (Level 1), N_DE=20 (Level 2) Output: Pareto front of optimized architectures // === LEVEL 1: Macro-Architecture Search (Aging Evolution) === Initialize population Pop = {A_1, ..., A_P} with random architectures from A history = [] FOR generation g = 1 TO 200 DO sample = RandomSample(Pop, size=5) // Tournament selection parent = argmax_{A in sample} ProxyFitness(A) child = Mutate(parent) // Single mutation child.fitness = ProxyFitness(child, D) // 20-epoch, 2-session LOSO Pop = Pop ∪ {child} Remove oldest individual from Pop // Aging mechanism history.append(child) END FOR TopK = SelectTopK(history, K=5) by proxy fitness // === LEVEL 2: Micro-Architecture Optimization (Differential Evolution) === ParetoFront = {} FOR EACH A_macro IN TopK DO Initialize DE population Θ = {θ_1, ..., θ_N_DE} randomly FOR generation g = 1 TO 50 DO FOR i = 1 TO N_DE DO Select r1, r2, r3 ∈ {1,...,N_DE} \ {i} randomly v_i = θ_r1 + F * (θ_r2 - θ_r3) // DE mutation u_i = BinomialCrossover(θ_i, v_i, CR) // Crossover IF FullFitness(A_macro, u_i, D) > FullFitness(A_macro, θ_i, D) THEN θ_i = u_i // Selection END IF END FOR END FOR θ* = argmax_{θ ∈ Θ} FullFitness(A_macro, θ, D) A_optimized = Configure(A_macro, θ*) ParetoFront = UpdateParetoFront(ParetoFront, A_optimized) END FOR RETURN ParetoFront // Non-dominated architectures: (Acc, |A|, Latency)

Algorithm 2. ProxyFitness cross-session proxy evaluation.

Input: Architecture A, Dataset D with S sessions, proxy_epochs=20 Output: Proxy fitness score (cross-session accuracy) // Use 2 randomly selected sessions for proxy evaluation s1, s2 = RandomSelect(sessions, 2) acc_list = [] // Fold 1: Train on s1, test on s2 model_1 = BuildModel(A) Train(model_1, D^{s1}, epochs=proxy_epochs) acc_1 = Evaluate(model_1, D^{s2}) acc_list.append(acc_1) // Fold 2: Train on s2, test on s1 model_2 = BuildModel(A) Train(model_2, D^{s2}, epochs=proxy_epochs) acc_2 = Evaluate(model_2, D^{s1}) acc_list.append(acc_2) RETURN mean(acc_list)

3.4 | Session-Invariant Feature Alignment Module

The SIFAM is a key architectural component that performs unsupervised domain adaptation between EEG sessions within the network. Unlike post-hoc domain adaptation methods that are applied independently of the classification architecture, SIFAM is integrated within the candidate architecture, and its internal structure is jointly optimized by the NAS alongside the main classifier, enabling the search process to discover the optimal configuration for aligning session-specific distributions while preserving class-discriminative information.

SIFAM operates on intermediate feature representations. During training, let $F_{src} = \{f_1^{src}, \ldots, f_n^{src}\}$ denote the features extracted from source session(s) at the SIFAM insertion point, and $F_{tgt} = \{f_1^{tgt}, \ldots, f_m^{tgt}\}$ denote features from the target session (available as unlabeled data during training). SIFAM minimizes the distributional discrepancy between these feature sets using the MMD criterion:

$$L_{SIFAM} = \lambda_{MMD} \cdot MMD^2(F_{src}, F_{tgt}), \quad (3)$$

where the squared MMD is computed as:

$$MMD^2(P, Q) = (1/n^2)\, \Sigma_{i,j}\, k(p_i, p_j) - (2/nm)\, \Sigma_{i,j}\, k(p_i, q_j) + (1/m^2)\, \Sigma_{i,j}\, k(q_i, q_j). \quad (4)$$

The kernel function k is selected from three options as part of the NAS search: 1) a linear kernel $k(x, y) = x^T y$, 2) a Gaussian RBF kernel $k(x, y) = \exp(-||x - y||^2 / 2\sigma^2)$ with bandwidth σ determined by the median heuristic (σ = median of pairwise distances), and 3) a multi-kernel combination:

$$k_{multi}(x, y) = \Sigma_b\, \alpha_b \cdot k_{RBF}(x, y; \sigma_b). \quad (5)$$

With bandwidths σ_b selected as multiples of the median heuristic ($\{0.5\sigma, \sigma, 2\sigma\}$) and uniform weights $\alpha_b = 1/3$. The total training loss combines classification and alignment objectives:

$$L_{total} = L_{classification} + L_{SIFAM}, \quad (6)$$

where $L_{classification}$ is the standard categorical cross-entropy loss. The alignment strength λ_{MMD} is searchable during the NAS process from {0.01, 0.05, 0.1, 0.5, 1.0}, as is the number of alignment layers within SIFAM (1, 2, or 3 fully connected layers that transform features before MMD computation) and the position of SIFAM within the main architecture (after cell 2, 4, or 6).

By making SIFAM's configuration part of the architecture search, MetaNAS-SI discovers the optimal domain adaptation strategy jointly with the classification architecture, a capability absent from all prior work that treats domain adaptation as a separate, manually configured post-processing step.

3.5 | Electroencephalography Data Preprocessing and Augmentation

All EEG data undergo a standardized preprocessing pipeline before entering the network. Continuous EEG recordings are band-pass filtered between 0.5 and 45 Hz using a 4th-order zero-phase Butterworth filter to remove DC drift and high-frequency noise while preserving the physiologically relevant EEG frequency bands. A Common Average Reference (CAR) is applied to reduce common-mode noise. Epochs are extracted from −0.5 to 2.0 seconds relative to the Speech Imagery cue onset, yielding 640-sample epochs at 256 Hz. Baseline correction is performed by subtracting the mean amplitude of the pre-stimulus period (−0.5 to 0 seconds) from each channel. Trials with peak-to-peak amplitude exceeding 100 μV on any channel are rejected as artifacts. Finally, z-score normalization is applied independently per channel per session to partially reduce session-level amplitude differences.

Data augmentation is applied during training to increase the effective training set size and improve generalization. Four augmentation strategies are employed, with their respective strengths optimized as continuous hyperparameters by the DE search in *Level 2*:

I. Gaussian noise injection: additive Gaussian noise with standard deviation σ_{noise}, searched in the range [0.01, 0.1], is added to each trial independently.

II. Temporal jittering: random temporal shifts of ±5 samples (approximately ±20 ms) are applied to each trial, simulating slight variations in imagery onset timing.

III. Channel dropout: 1–3 channels are randomly zeroed in each trial during training, encouraging the network to be robust to individual channel noise or loss.

IV. Sliding window cropping: random temporal crops of 80–100% of the trial length are extracted and zero-padded to the original length, providing temporal variability.

3.6 | Pareto-Optimal Architecture Selection

After the two-level NAS completes, the set of evaluated architectures yields a Pareto front of non-dominated solutions across three objectives: cross-session LOSO accuracy, model parameter count, and inference latency. An architecture is Pareto-optimal (non-dominated) if no other architecture is simultaneously better on all three objectives. The Pareto front provides a menu of architectures from which practitioners can select based on deployment requirements:

I. Maximum accuracy: the architecture with the highest cross-session accuracy, suitable for offline analysis applications where computational resources are unconstrained.

II. Balanced (knee-point): the architecture at the knee point of the Pareto front, defined as the point with minimum normalized Euclidean distance to the ideal point (maximum accuracy, minimum parameters, minimum latency). Such a selection represents the best trade-off for general-purpose BCI applications.

III. Minimum latency: the architecture with the lowest inference latency that still exceeds a minimum accuracy threshold (chance level + 10 percentage points). Such an architecture is suitable for real-time BCI applications requiring <50 ms inference.

The Pareto-optimal selection is performed using the non-dominated sorting procedure from NSGA-II [55], applied as a post-processing step after all architecture evaluations are complete.

3.7 | Computational Cost Analysis

The total computational cost of MetaNAS-SI is summarized as follows:

Table 2. Computational cost analysis of the MetaNAS-SI framework.

Phase	Configuration	Cost Per Evaluation	Total Cost
Level 1 (aging evolution)	P=25, 200 generations, ~200 architectures	~15 min (20 epochs, 2-session proxy)	~50 GPU-hours
Level 2 (DE)	5 architectures × 20 pop × 50 gen	~45 min (100 epochs, full LOSO)	~40 GPU-hours
Total NAS			~90 GPU-hours
Final training (per subject)	100 epochs, full data		~1.5 hours

All experiments are conducted on an NVIDIA RTX 3090 GPU (24 GB VRAM). The 90 GPU-hour NAS cost is a one-time investment per dataset and Speech Imagery paradigm. Once the optimal architecture is discovered, it is fixed and can be trained for new subjects in approximately 1.5 hours, making MetaNAS-SI practical for clinical BCI deployment.

4 | Experimental Setup

4.1 | Datasets

MetaNAS-SI was evaluated on three Speech Imagery EEG datasets spanning different languages, channel configurations, and class counts to assess generalization across diverse experimental conditions.

4.1.1 | KaraOne dataset

The KaraOne dataset [21] comprises EEG recordings from 12 subjects performing imagined speech of 7 phonemes and words: /iy/, /uw/, /piy/, /tiy/, /diy/, /m/, and /n/, plus a rest condition. Recordings were made using a 62-channel EEG system at a sampling rate of 256 Hz. Each class was repeated 12 times per run, and each subject completed 3 sessions on different days. Following prior work [24], [26], we use a 5-class subset consisting of the phonemes /iy/, /uw/, /piy/, /tiy/, and /diy/ to enable comparison with published baselines.

4.1.2 | Coretto imagined speech dataset

The Coretto Imagined Speech dataset [22] contains EEG recordings from 15 subjects imagining the production of 4 Spanish vowels: /a/, /e/, /i/, and /u/. Recordings were acquired using a 14-channel Emotiv EPOC headset at 1024 Hz (downsampled to 256 Hz for our experiments). Each subject completed 40 trials per class per session across 3 sessions. The lower channel count (14 channels) presents an additional challenge for spatial feature extraction compared to the 62-channel KaraOne dataset.

4.1.3 | Farsi-Arabic imagined speech dataset

The FA-IS dataset is an in-house dataset collected specifically for this study. Twelve subjects (6 native Farsi speakers, 6 native Arabic speakers; 7 male, 5 female; age range 22–35 years) were recorded using a 64-channel BioSemi ActiveTwo EEG system at 256 Hz. Subjects imagined speaking 6 words: 3 Farsi words (/salam/ [hello], /bale/ [yes], /na/ [no]) and 3 Arabic words (/marhaba/ [hello], /naam/ [yes], /la/ [no]). Each subject completed 30 trials per class per session across 5 sessions, with sessions spread over approximately 2 weeks (2–4 days between sessions). The 5-session design provides richer cross-session variability data compared to the 3-session KaraOne and Coretto datasets, enabling more robust LOSO evaluation and NAS fitness estimation. This study was approved by the Institutional Review Board of Sharif University of Technology (IRB-SUT-2024-CS-087) and the United Arab Emirates University Research Ethics Committee (UAEU-REC-2024-142). All participants provided written informed consent.

Table 3. Summary of datasets used for evaluation.

Property	KaraOne	Coretto	FA-IS
Subjects	12	15	12
EEG channels	62	14	64
Sampling rate	256 Hz	1024 Hz (resampled to 256 Hz)	256 Hz
Classes	5 (phonemes)	4 (vowels)	6 (words)
Sessions per subject	3	3	5
Trials per class per session	12	40	30
Language	English	Spanish	Farsi / Arabic

4.2 | Evaluation Protocol

Two evaluation protocols are employed, with cross-session LOSO serving as the primary metric:

I. Within-session (baseline comparison): 5-fold stratified cross-validation within each session. Performance is averaged across folds and sessions. Such an evaluation serves as an upper-bound reference representing the best-case scenario without session transfer.

II. Cross-session LOSO (primary metric): LOSO evaluation. For a subject with S sessions, the model is trained on $S-1$ sessions and tested on the held-out session. This is repeated for each session fold, and accuracy is averaged across all folds and subjects. This protocol directly measures cross-session generalization capability.

III. Cross-subject LOSO (secondary analysis): leave-one-subject-out evaluation. The model is trained on data from all subjects except one and tested on the held-out subject. Such an evaluation assesses cross-subject generalization, which is a more challenging transfer scenario.

We report the following metrics: classification accuracy (mean ± standard deviation across subjects), Cohen's kappa coefficient (κ), macro-averaged F1-score, and per-class confusion matrices for the best-performing architecture.

4.3 | Compared Methods

MetaNAS-SI is compared against 10 baseline methods spanning traditional machine learning, hand-designed deep learning, and transfer learning approaches:

4.3.1 | Traditional machine learning

I. CSP + LDA: CSP features (6 filter pairs) classified by linear discriminant analysis [56]. One-vs-rest extension for multi-class classification.

II. FBCSP + SVM: FBCSP with 9 frequency sub-bands (4–40 Hz, 4 Hz bandwidth) and mutual information-based feature selection, classified by SVM with RBF kernel [57].

III. PSD + RF: PSD features extracted via Welch's method in theta, alpha, beta, and gamma bands, classified by a Random Forest with 500 trees.

4.3.2 | Hand-designed deep learning

I. EEGNet [13]: compact CNN with depthwise and separable convolutions, using the recommended configuration (F1=8, D=2, F2=16, kernel length=64).

II. DeepConvNet [14]: a deeper CNN architecture with 4 convolutional-max pooling blocks designed for EEG raw signal classification.

III. ShallowConvNet [14]: a shallower architecture inspired by FBCSP that uses a temporal convolution followed by a spatial filter and a squaring nonlinearity.

IV. EEG-conformer [15]: a hybrid CNN-transformer architecture that combines convolutional feature extraction with self-attention for temporal modeling of EEG signals.

V. EEG-Temporal Convolutional Network (TCNet): TCNet for EEG incorporating dilated causal convolutions for long-range temporal dependency modeling [58].

4.3.3 | Transfer learning baselines

I. EEGNet + EA: EEGNet with EA [28] applied to align the covariance matrices of training and test sessions before classification.

II. EEGNet + fine-tune: EEGNet pre-trained on source session data and fine-tuned on a small amount of labeled data from the target session (10% of target session trials used for fine-tuning).

4.4 | Implementation Details

All models are implemented in PyTorch 2.0 and trained on an NVIDIA RTX 3090 GPU. MetaNAS-SI search parameters: Aging evolution population P=25, tournament size S=5, 200 generations, proxy training=20 epochs. DE parameters: N_{DE}=20, F=0.5, CR=0.9, 50 generations, full training=100 epochs. Training uses the Adam optimizer [59] with learning rate, batch size, and weight decay optimized by DE from the ranges $\{1\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}\}$, {16, 32, 64}, and $\{1\times10^{-4}, 1\times10^{-3}, 1\times10^{-2}\}$, respectively. Early stopping with patience of 15 epochs is applied during all training runs. SIFAM MMD kernel bandwidths are set using the median heuristic computed on the training data. All baseline methods are reproduced using publicly available implementations with hyperparameters optimized via grid search on a validation split to ensure fair comparison.

4.5 | Statistical Analysis

Statistical significance of accuracy differences between MetaNAS-SI and each baseline is assessed using the paired Wilcoxon signed-rank test across subjects at a significance level of $\alpha = 0.05$, with Bonferroni correction for multiple comparisons (10 comparisons). Effect size is quantified using Cohen's d. Significance levels are denoted as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5 | Results

5.1 | Cross-Session Results on KaraOne (5-Class)

Table 4 presents the cross-session LOSO classification accuracy for all methods on the KaraOne dataset (5-class, 12 subjects, 3 sessions per subject). MetaNAS-SI achieved a mean cross-session accuracy of 45.8 ± 4.2%, substantially outperforming all 10 baseline methods. The best baseline, EEGNet + EA, achieved 36.4 ± 4.8%, yielding a 9.4 percentage-point absolute improvement (25.8% relative improvement). The improvement over the standard EEGNet baseline was 14.6 percentage points (46.8% relative). All pairwise comparisons between MetaNAS-SI and baselines were statistically significant at $p < 0.001$ after Bonferroni correction, with large effect sizes (Cohen's $d > 1.5$ for all comparisons).

Table 4. Cross-session LOSO classification accuracy (%) on KaraOne (5-class, chance = 20%). Mean ± standard deviation across 12 subjects. κ: Cohen's kappa. F1: macro-averaged F1-score.

Method	Accuracy (%)	κ	F1 (%)	Params	Sig.
CSP + LDA	28.5 ± 6.2	0.106	26.3		***
FBCSP + SVM	30.1 ± 5.7	0.126	28.5		***
PSD + RF	27.3 ± 5.5	0.091	25.1		***
EEGNet	31.2 ± 5.1	0.140	29.4	~170K	***
DeepConvNet	29.8 ± 5.8	0.123	27.9	~2.1M	***
ShallowConvNet	28.4 ± 4.9	0.105	26.8	~47K	***
EEG-conformer	34.1 ± 4.6	0.176	32.3	~350K	***
EEG-TCNet	32.7 ± 5.3	0.159	30.8	~4.1K	***
EEGNet + EA	36.4 ± 4.8	0.205	34.6	~170K	***
EEGNet + fine-tune	33.5 ± 5.0	0.169	31.7	~170K	***
MetaNAS-SI (ours)	45.8 ± 4.2	0.323	43.9	~42K	

Among the deep learning baselines, EEG-Conformer (34.1%) outperformed EEGNet (31.2%), suggesting that the attention mechanism captures temporal dependencies relevant to Speech Imagery. However, the improvement from EEG-Conformer over EEGNet (+2.9 pp) was modest compared to MetaNAS-SI's

improvement over EEG-Conformer (+11.7 pp), indicating that architectural innovation alone is insufficient without cross-session optimization. Traditional methods (CSP+LDA, FBCSP+SVM, PSD+RF) performed near or below 30%, consistent with the known limitations of hand-crafted features for the weak and diffuse signals of Speech Imagery.

The within-session comparison (*Table 5*) reveals a complementary pattern. MetaNAS-SI achieved 62.4 ± 5.1% within-session accuracy, compared to 56.8 ± 5.7% for EEG-Conformer and 54.3 ± 6.2% for EEGNet. The within-session improvement (5.6–8.1 pp) is notably smaller than the cross-session improvement (9.4–14.6 pp), confirming that MetaNAS-SI's cross-session fitness function specifically optimizes for session-transfer robustness rather than merely improving general classification capability.

Table 5. Within-session accuracy (%) on KaraOne (5-class, 5-fold CV). Mean ± SD across 12 subjects.

Method	Within-Session Accuracy (%)	Cross-Session Accuracy (%)	Drop (%)
EEGNet	54.3 ± 6.2	31.2 ± 5.1	−23.1
EEG-Conformer	56.8 ± 5.7	34.1 ± 4.6	−22.7
EEGNet + EA	54.3 ± 6.2	36.4 ± 4.8	−17.9
MetaNAS-SI	62.4 ± 5.1	45.8 ± 4.2	−16.6

MetaNAS-SI also exhibits the smallest within-to-cross-session accuracy drop (−16.6 pp) compared to EEGNet (−23.1 pp) and EEG-conformer (−22.7 pp), demonstrating superior session-transfer robustness. EEGNet + EA reduces the drop to −17.9 pp through EA, approaching MetaNAS-SI's robustness, but at substantially lower absolute accuracy.

5.2 | Cross-Session Results on Coretto (4-Class)

Table 6 presents the cross-session LOSO results on the Coretto imagined speech dataset (4-class vowel classification, 15 subjects). MetaNAS-SI achieved 52.3 ± 3.8%, compared to 43.8 ± 3.9% for the best baseline (EEGNet + EA), representing an 8.5 percentage-point improvement (19.4% relative). The higher absolute accuracies across all methods on Coretto compared to KaraOne are expected given the fewer classes (4 vs. 5) and the phonological distinctiveness of the four vowels (/a/, /e/, /i/, /u/).

Table 6. Cross-session LOSO accuracy (%) on Coretto imagined speech (4-class, chance = 25%). Mean ± SD across 15 subjects.

Method	Accuracy (%)	κ	F1 (%)
CSP + LDA	32.1 ± 5.4	0.095	30.2
FBCSP + SVM	34.5 ± 5.1	0.127	32.8
PSD + RF	31.4 ± 4.8	0.085	29.5
EEGNet	38.7 ± 4.5	0.183	36.9
DeepConvNet	36.2 ± 5.2	0.149	34.1
ShallowConvNet	35.8 ± 4.7	0.144	33.9
EEG-conformer	41.2 ± 4.1	0.216	39.4
EEG-TCNet	39.5 ± 4.6	0.193	37.6
EEGNet + EA	43.8 ± 3.9	0.251	41.7
EEGNet + fine-tune	40.6 ± 4.3	0.208	38.5
MetaNAS-SI (ours)	52.3 ± 3.8	0.364	50.6

Notably, MetaNAS-SI achieves a Cohen's kappa of 0.364, substantially above the 0.251 of EEGNet + EA, indicating better agreement between predictions and true labels beyond chance. The lower standard deviation (3.8%) across 15 subjects also suggests that MetaNAS-SI provides more consistent performance across individuals.

5.3 | Cross-Session Results on Farsi-Arabic Imagined Speech (6-Class, 5 Sessions)

Table 7 presents results on the FA-IS dataset, the most challenging evaluation due to the higher class count (6 classes, chance = 16.7%) and the longest session span (5 sessions over 2 weeks). MetaNAS-SI achieved 48.1

± 4.5%, compared to 35.2 ± 4.6% for the best baseline (EEGNet + EA), a 12.9 percentage-point absolute improvement (36.6% relative).

Table 7. Cross-session LOSO accuracy (%) on FA-IS (6-class, chance = 16.7%). Mean ± SD across 12 subjects.

Method	Accuracy (%)	κ	F1 (%)
CSP + LDA	22.4 ± 5.9	0.069	20.1
FBCSP + SVM	24.8 ± 5.5	0.098	22.7
PSD + RF	21.7 ± 5.3	0.060	19.5
EEGNet	30.5 ± 5.2	0.166	28.3
DeepConvNet	28.1 ± 5.6	0.137	25.9
ShallowConvNet	27.3 ± 5.0	0.128	25.2
EEG-Conformer	33.8 ± 4.8	0.206	31.5
EEG-TCNet	31.6 ± 5.1	0.179	29.4
EEGNet + EA	35.2 ± 4.6	0.222	33.1
EEGNet + Fine-tune	32.1 ± 5.0	0.185	29.8
MetaNAS-SI (Ours)	48.1 ± 4.5	0.377	46.2

The larger improvement on FA-IS compared to KaraOne and Coretto can be attributed to the 5-session design providing more cross-session variability for the NAS fitness function to exploit. With more session pairs available during architecture evaluation, MetaNAS-SI can more reliably identify architectures that generalize across session-specific confounds. Analysis by native language revealed comparable performance: MetaNAS-SI achieved 49.3 ± 4.1% for Farsi speakers and 46.9 ± 4.9% for Arabic speakers. The difference was not statistically significant ($p = 0.31$, Wilcoxon), suggesting that the discovered architecture captures language-invariant Speech Imagery representations.

5.4 | Discovered Architecture Analysis

Analysis of the Pareto-optimal architectures discovered by MetaNAS-SI reveals several notable patterns. *Table 8* summarizes the three representative architectures from the Pareto front, corresponding to maximum accuracy, balanced, and minimum latency configurations.

Table 8. Characteristics of Pareto-optimal architectures discovered by MetaNAS-SI on KaraOne.

Architecture	Cells	Key Operations	SIFAM Configuration	Params	Accuracy (%)	Latency (ms)
A (max accuracy)	6	TC-multi, SC-CSP, TA-self (4 heads)	Cell 4, 2 layers, multi-kernel	42K	45.8	12
B (balanced)	4	TC-Alpha, TC-Beta, SC-attention, TA-relative	Cell 2, 1 layer, RBF	18K	43.2	8
C (min latency)	3	TC-Beta, SC-full	Cell 2, 1 layer, linear	8K	39.7	3

Several observations emerge from the discovered architectures:

I. Multi-scale temporal convolutions are consistently preferred. The highest-accuracy architecture (A) selects TC-Multi (parallel multi-scale convolution), while the balanced architecture (B) selects two specific temporal convolutions (TC-Alpha and TC-Beta). These findings confirm the neuroscientific understanding that Speech Imagery engages multiple frequency bands simultaneously and that architectures capturing this multi-scale temporal structure outperform those restricted to a single temporal resolution.

II. CSP-inspired spatial filters outperform standard approaches. Architecture A selects SC-CSP over SC-Full, demonstrating that incorporating prior knowledge from the decades-long CSP tradition in BCI engineering into the learnable architecture yields better spatial feature extraction. The SC-CSP operation effectively bridges classical BCI signal processing with modern deep learning.

III. SIFAM is optimally placed in middle layers. Across all Pareto-optimal architectures, SIFAM is placed after cell 2 or cell 4, never at the earliest or latest cells. These findings suggest that early layers should retain session-specific low-level information (which SIFAM would prematurely remove), while late layers should produce session-invariant, class-discriminative representations (which SIFAM has already facilitated in the middle layers).

IV. Discovered architectures are remarkably compact. With parameter counts ranging from 8K to 42K, the MetaNAS-SI architectures are 4–50× smaller than EEGNet (~170K parameters) and 50–260× smaller than DeepConvNet (~2.1M parameters), while achieving substantially higher cross-session accuracy. This compactness suggests that Speech Imagery decoding does not benefit from overparameterized architectures but rather requires the right inductive biases, which MetaNAS-SI discovers automatically.

5.5 | Cross-Subject Generalization

While cross-session transfer is the primary focus of this work, we also assessed cross-subject generalization using the more challenging leave-one-subject-out protocol. *Table 9* presents results.

Table 9. Cross-subject LOSO accuracy (%) on all three datasets. Mean ± SD.

Method	KaraOne (5-Class)	Coretto (4-Class)	FA-IS (6-Class)
EEGNet	24.1 ± 6.2	29.5 ± 5.0	20.8 ± 5.4
EEG-Conformer	26.3 ± 5.7	31.2 ± 4.6	22.4 ± 5.1
EEGNet + EA	28.5 ± 5.4	34.1 ± 4.3	24.7 ± 4.9
MetaNAS-SI	33.2 ± 5.8	39.1 ± 4.2	31.8 ± 5.1

Cross-subject accuracy is substantially lower than cross-session accuracy for all methods, reflecting the greater distributional shift between subjects (anatomical differences, individual cognitive strategies, neurophysiological variability) compared to between sessions of the same subject. Nevertheless, MetaNAS-SI achieves the largest improvement in cross-subject settings as well (4.7–7.1 pp over best baseline), suggesting that SIFAM and the cross-session fitness function also confer partial robustness to subject-level distribution shifts.

5.6 | Ablation Study

To quantify the contribution of each component, we conducted a systematic ablation study on the KaraOne dataset, removing one component at a time from the full MetaNAS-SI framework. Results are presented in *Table 10*.

Table 10. Ablation study on KaraOne cross-session LOSO accuracy (%). Each row removes one component from the full MetaNAS-SI framework.

Configuration	Accuracy (%)	Δ (pp)
Full MetaNAS-SI	45.8	
− Cross-session fitness (within-session optimization)	35.2	−10.6
− SIFAM (no domain adaptation)	38.4	−7.4
− Aging evolution (random architecture sampling)	37.8	−8.0
− EEG-specific search space (generic operations)	40.1	−5.7
− Data augmentation	41.6	−4.2
− DE micro-optimization (default hyperparameters)	42.3	−3.5
− Multi-objective (accuracy only, no size constraint)	46.1	+0.3

The two most impactful components are the cross-session fitness function (−10.6 pp when removed) and the aging evolution search algorithm (−8.0 pp when replaced with random search). The cross-session fitness function is the single most important innovation: Without it, the NAS discovers architectures that overfit to within-session patterns, and the resulting architecture achieves only 35.2% cross-session accuracy, comparable to hand-designed baselines. These findings confirm our central hypothesis that explicitly optimizing for cross-session generalization during architecture search is critical.

Removing SIFAM (−7.4 pp) demonstrates that domain adaptation is a crucial complementary mechanism: even with cross-session fitness, the absence of explicit feature alignment substantially degrades performance. The EEG-specific search space contributes −5.7 pp; replacing it with generic operations (standard convolutions with arbitrary kernel sizes, no CSP-inspired filters) yields an architecture that fails to exploit the oscillatory structure of EEG signals. Data augmentation (−4.2 pp) and DE micro-optimization (−3.5 pp) provide smaller but meaningful contributions.

Interestingly, removing the multi-objective constraint (allowing the search to optimize only for accuracy without penalizing model size) yields a marginal accuracy increase (+0.3 pp) but produces a model 5× larger (~210K parameters). These findings confirm that the Pareto-optimal architecture selection imposes minimal accuracy cost while dramatically reducing model size.

5.7 | Session-by-Session Analysis

To examine the consistency of cross-session transfer across different session pairs, *Table 11* presents a per-session breakdown for a representative subject (subject 3) from the KaraOne dataset.

Table 11. Per-session pair accuracy (%) for Subject 3 on KaraOne. Si→Sj denotes training on session i and testing on session j.

Method	S1→S2	S1→S3	S2→S3	Mean	SD
EEGNet	32.4	27.8	33.1	31.1	2.9
EEGNet + EA	38.2	33.5	39.1	36.9	3.0
MetaNAS-SI	47.1	43.5	48.2	46.3	2.5

MetaNAS-SI achieves not only higher mean accuracy but also lower variance across session pairs (SD = 2.5 vs. 2.9–3.0 for baselines), indicating more consistent and reliable cross-session transfer. The S1→S3 transfer (largest temporal gap between sessions) is the most challenging for all methods, but MetaNAS-SI maintains a relatively small drop compared to adjacent-session transfers.

5.8 | Computational Cost

Table 12 summarizes the computational characteristics of MetaNAS-SI and baselines.

Table 12. Computational cost comparison. NAS cost is a one-time investment; training and inference costs are per-subject.

Method	Architecture Design	Training Time	Inference Latency	Parameters
EEGNet	Manual (months of expert effort)	~0.5 h	5 ms	~170K
DeepConvNet	Manual	~1.2 h	15 ms	~2.1M
EEG-conformer	Manual	~2.0 h	22 ms	~350K
MetaNAS-SI (A)	~90 GPU-h (automated, one-time)	~1.5 h	12 ms	~42K
MetaNAS-SI (B)	~90 GPU-h (automated, one-time)	~1.0 h	8 ms	~18K
MetaNAS-SI (C)	~90 GPU-h (automated, one-time)	~0.4 h	3 ms	~8K

All MetaNAS-SI architectures achieve inference latency well below the 50 ms threshold required for real-time BCI operation at a 256 Hz sampling rate. The minimum-latency architecture (C) achieves 3 ms inference, enabling deployment on edge devices with limited computational resources. While the one-time NAS search cost of ~90 GPU-hours is substantial, it is amortized across all future subjects and sessions and replaces months of manual architecture engineering effort.

5.9 | Frequency Band Importance Analysis

To provide neurophysiological interpretability, we analyzed the contribution of each temporal convolution kernel in the discovered Architecture A using gradient-based feature importance (integrated gradients [60]). *Table 13* presents the relative importance of each temporal convolution operation.

Table 13. Frequency band importance in MetaNAS-SI Architecture A, measured by integrated gradient feature importance (normalized to sum to 100%).

Temporal Operation	Target Band	Frequency Range (Hz)	Importance (%)
TC-Beta	Beta	13–30	32.1
TC-Alpha	Alpha	8–13	24.5
TC-Gamma	Gamma	30–100	18.7
TC-Theta	Theta	4–8	15.2
TC-Delta	Delta	0.5–4	9.5

The beta band (13–30 Hz) emerged as the most important frequency range, contributing 32.1% of the total feature importance. This finding is consistent with the well-established role of beta oscillations in motor planning and speech production: Beta desynchronization in sensorimotor and frontal regions accompanies both overt and imagined speech [49], [61]. The alpha band (8–13 Hz, 24.5% importance) likely reflects alpha event-related desynchronization associated with cortical activation during imagery. The gamma band (30–100 Hz, 18.7% importance) is linked to high-level language processing, with gamma synchronization in Broca's and Wernicke's areas during language tasks [50]. The theta band (4–8 Hz, 15.2% importance) reflects working memory and semantic processing demands inherent to imagined speech [51]. The delta band contributes the least (9.5%), consistent with its primary association with deep sleep and autonomic processes rather than active cognitive tasks. These findings demonstrate that MetaNAS-SI discovers architectures whose learned features align with established neuroscientific understanding of speech-related cortical oscillations, enhancing the interpretability and trustworthiness of the discovered architectures.

6 | Discussion

The results presented in this study demonstrate that MetaNAS-SI achieves substantial improvements in cross-session Speech Imagery EEG decoding across three datasets, two languages, and varying channel configurations. Here we discuss the mechanisms underlying these improvements, their implications for BCI research and clinical practice, and the limitations of the current work.

6.1 | Why Cross-Session Fitness is Transformative

The ablation study reveals that the cross-session fitness function is the single most impactful component of MetaNAS-SI, contributing 10.6 percentage points to cross-session accuracy on KaraOne. Standard NAS approaches optimize within-session accuracy, inadvertently selecting architectures that exploit session-specific confounding patterns in the EEG that are consistent within a session but do not transfer across sessions. These confounds include session-specific electrode impedance profiles that create characteristic spatial noise patterns, environmental electromagnetic interference that differs between recording days, and the subject's momentary cognitive strategy or arousal level. By evaluating candidate architectures using a LOSO protocol, MetaNAS-SI applies selection pressure that directly penalizes reliance on such non-transferable patterns and rewards extraction of session-invariant neural features. This simple but powerful shift in the optimization target fundamentally changes which architectures are discovered, favoring those with the right inductive biases for cross-session robustness.

6.2 | Session-Invariant Feature Alignment Module's Mechanism and Optimal Placement

SIFAM contributes 7.4% points to cross-session accuracy by explicitly minimizing the distributional distance between session-specific feature representations at an intermediate network layer. By forcing the MMD between source and target session features toward zero, SIFAM encourages the network to learn representations that are invariant to session-specific confounds while preserving class-discriminative information (because the classification loss is optimized simultaneously). The finding that SIFAM is optimally placed in middle layers (cell 2–4) rather than at the input or output has an intuitive interpretation: Early layers should retain session-specific low-level information (e.g., signal amplitude scaling, channel-specific noise patterns) that is useful for the subsequent alignment process, while late layers should already operate on session-invariant representations. Placing SIFAM too early would force alignment of raw signals before meaningful features are extracted, while placing it too late would miss the opportunity to guide the network toward invariant intermediate representations.

6.3 | Electroencephalography-Specific Search Space Advantages

The 5.7% point improvement contributed by the EEG-specific search space (compared to generic operations) validates the hypothesis that constraining the NAS to neurophysiologically plausible operations yields better

architectures more efficiently. The physiologically-motivated temporal convolution kernels ensure that the network can efficiently extract features from canonical EEG frequency bands without needing to learn appropriate filter lengths from scratch. The CSP-inspired spatial filters embed decades of BCI engineering knowledge into the learnable architecture, providing a structured inductive bias that standard spatial convolutions lack. This approach represents a productive middle ground between fully hand-designed architectures (which may miss novel architectural possibilities) and unconstrained NAS (which wastes search budget exploring neurophysiologically implausible operations).

6.4 | Discovered Architecture Insights

The remarkably compact size of discovered architectures (8K–42K parameters vs. 170K–2.1M for baselines) challenges the assumption that larger models are needed for EEG decoding. EEG signals, particularly from Speech Imagery, have relatively limited information content compared to natural images or audio, and the small training datasets typical in BCI (tens to hundreds of trials per class) cannot support highly overparameterized models without severe overfitting. MetaNAS-SI discovers that the critical ingredients for Speech Imagery decoding are multi-scale temporal feature extraction, spatially informed filtering, and domain adaptation, not model depth or width. This finding has practical implications: Compact models are faster to train, require less memory, and can be deployed on embedded BCI hardware.

6.5 | Clinical Implications

The 45.8% accuracy for 5-class imagined speech classification in cross-session evaluation on KaraOne represents, to our knowledge, the highest reported cross-session performance on this dataset. While this accuracy is not yet sufficient for reliable communication (which typically requires >70% accuracy for practical BCI communication systems [1]), it represents a meaningful step toward clinical viability. The cross-session robustness of MetaNAS-SI is particularly relevant for clinical deployment, where daily recalibration is burdensome for patients with severe motor impairments. Combining MetaNAS-SI with improved recording protocols (e.g., standardized electrode placement using 3D digitization), larger training datasets (enabled by multi-day recordings), and complementary modalities (e.g., Functional Near-Infrared Spectroscopy (fNIRS)) could push accuracy toward clinically useful levels. The Pareto-optimal architecture selection also enables clinicians to choose deployment configurations appropriate for the clinical context: Maximum accuracy for research-grade offline analysis, balanced for general-purpose BCI use, or minimum latency for real-time feedback applications.

6.6 | Limitations

Several limitations of the current work should be acknowledged. First, the NAS search cost of approximately 90 GPU-hours, while a one-time investment, may be prohibitive for researchers without access to GPU computing infrastructure. Transfer of discovered architectures across datasets could reduce this cost but requires further validation. Second, evaluation was conducted on relatively small datasets (12–15 subjects, 3–5 sessions), and the generalizability of the discovered architectures to larger populations and longer-term studies remains to be established. Third, the search space, while EEG-specific, is manually designed based on our domain knowledge; automatic search space generation through meta-learning could further improve results. Fourth, MetaNAS-SI discovers a fixed architecture after the search process; online adaptation of the architecture during BCI use (e.g., adapting to changing neural states within a session) is not addressed. Fifth, only discrete imagined words and phonemes were evaluated; continuous imagined speech decoding, which is a substantially harder problem requiring temporal sequence models, is beyond the scope of this work. Sixth, the current framework uses EEG as the sole input modality; multimodal fusion with complementary signals such as fNIRS or Electromyography (EMG) from articulatory muscles could provide additional discriminative information.

6.7 | Comparison with Related Neural Architecture Search for Electroencephalography

The limited existing NAS work in BCI [19], AutoEEG by Zhu et al. [20], focuses exclusively on motor imagery classification using within-subject evaluation. Neither incorporates cross-session objectives, searchable domain adaptation modules, or EEG-specific operations. MetaNAS-SI's cross-session fitness function and SIFAM module represent novel contributions to the NAS-BCI literature. Furthermore, the hybrid aging evolution + DE strategy addresses both the discrete and continuous aspects of the architecture optimization problem more naturally than purely differentiable or purely evolutionary approaches.

6.8 | Future Directions

Several promising research directions emerge from this work. Extension to online BCI with real-time architecture adaptation could enable the system to adjust its processing pipeline as the user's neural signals change within and across sessions. Transfer of MetaNAS-SI to intracranial EEG Electrocorticography (ECoG) could leverage higher signal quality for more accurate speech decoding while still benefiting from automated architecture optimization. Federated NAS across multiple BCI centers would enable architecture discovery from larger and more diverse datasets without centralizing sensitive neural data. Integration of meta-learning for few-shot session adaptation could reduce the calibration data required for a new session from hundreds of trials to tens. Continuous imagined speech decoding, where the user imagines speaking full sentences rather than isolated words, represents the ultimate goal for silent speech interfaces and could benefit from MetaNAS-SI's framework by incorporating sequence-to-sequence architectures into the search space. Finally, integration with large language models as post-processing could leverage linguistic priors to correct classification errors, effectively implementing a neural speech "autocorrect" system.

7 | Conclusion

This paper introduced MetaNAS-SI, a Metaheuristic-Guided Neural Architecture Search framework designed to address the critical cross-session transfer problem in EEG-based Speech Imagery decoding. MetaNAS-SI combines five synergistic innovations: A hybrid evolutionary search strategy coupling aging evolution for macro-architecture exploration with DE for micro-architecture optimization; a cross-session LOSO fitness function that directly optimizes for session-transfer robustness; a searchable SIFAM that integrates domain adaptation within the architecture search process; an EEG-specific search space incorporating neurophysiologically motivated temporal convolutions, CSP-inspired spatial filters, and multi-scale temporal attention; and Pareto-optimal architecture selection for deployment-aware trade-offs among accuracy, model size, and inference latency.

Evaluated on three Speech Imagery EEG datasets spanning English, Spanish, Farsi, and Arabic imagined speech, MetaNAS-SI achieved cross-session classification accuracies of 45.8% (KaraOne, 5-class), 52.3% (Coretto, 4-class), and 48.1% (FA-IS, 6-class), outperforming 10 baseline methods by 8.5–14.6 percentage points. The discovered architectures are 3–15× more compact than hand-designed alternatives, with inference latencies suitable for real-time BCI operation. Ablation analysis confirmed that the cross-session fitness function and SIFAM module are the two most impactful components, together contributing 18.0 percentage points to cross-session accuracy. Frequency band importance analysis revealed that the discovered architectures preferentially leverage beta, alpha, and gamma oscillations, consistent with established neuroscience of speech-related cortical activity, demonstrating both the effectiveness and interpretability of the automatically discovered architectures.

MetaNAS-SI represents a paradigm shift from manual, within-session-focused architecture design to automated, cross-session-optimized architecture discovery for Speech Imagery BCIs. While the achieved accuracy levels are not yet sufficient for clinical deployment, the substantial improvements over all baselines and the principled framework for cross-session optimization establish a foundation upon which future

advances in Speech Imagery BCI can build. The code, discovered architectures, and the FA-IS dataset will be made publicly available to facilitate reproducibility and future research.

Ethical Statement

This study was approved by the IRB-SUT and the UAEU-REC. All participants provided written informed consent prior to participation. The KaraOne and Coretto datasets are publicly available with appropriate ethical approvals from their original institutions. All data were stored securely and de-identified prior to analysis.

Acknowledgments

This research was supported by the Iran National Science Foundation (INSF) under Grant No. 4014112 and the UAE University UPAR Grant No. 12S167. The authors thank the participants of the FA-IS study for their valuable contribution and patience across multiple recording sessions.

Author Contributions

R. M.: conceptualization, methodology, software, writing–original draft, supervision, project administration. Al-F. S.: data curation, validation, formal analysis, writing–review and editing, visualization.

Funding

Not applicable.

Data Availability

The KaraOne dataset is available at http://www.cs.toronto.edu/karaone/. The Coretto Imagined Speech dataset is available at https://doi.org/10.5281/zenodo.7556647. The FA-IS dataset will be made available upon publication at https://github.com/rahimi-lab/MetaNAS-SI.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent for Publication

The author confirms consent for the publication of this work.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] McFarland, D. J., & Wolpaw, J. R. (2011). Brain-computer interfaces for communication and control. *Communications of the ACM*, *54*(5), 60–66. https://doi.org/10.1145/1941487.1941506

[2] Birbaumer, N., Murguialday, A. R., & Cohen, L. (2008). Brain-computer interface in paralysis. *Current opinion in neurology*, *21*(6), 634—638. https://doi.org/10.1097/wco.0b013e328315ee2d

[3] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of neural engineering*, *15*(3), 31005. https://doi.org/10.1088/1741-2552/aab2f2 1741-2552/aab2f2

[4] Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in cognitive sciences*, *8*(5), 204–210. https://doi.org/10.1016/j.tics.2004.03.008

[5] Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

[6] Martin, S., Iturrate, I., Millán, J. del R., Knight, R. T., & Pasley, B. N. (2018). Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience*, *12*, 1–10. https://doi.org/10.3389/fnins.2018.00422

[7] DaSalla, C. S., Kambara, H., Sato, M., & Koike, Y. (2009). Single-trial classification of vowel Speech Imagery using common spatial patterns. *Neural networks*, *22*(9), 1334–1339. https://doi.org/10.1016/j.neunet.2009.05.008

[8] Brigham, K., & Kumar, B. V. K. V. (2010). Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy. *2010 4th international conference on bioinformatics and biomedical engineering* (pp. 1–4). IEEE. https://doi.org/10.1109/ICBBE.2010.5515807

[9] Nguyen, C. H., Karavas, G. K., & Artemiadis, P. (2017). Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features. *Journal of neural engineering*, *15*(1), 16002. https://doi.org/10.1088/1741-2552/aa8235

[10] Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. N., & Müller, K. R. (2006). Towards adaptive classification for BCI. *Journal of neural engineering*, *3*(1), R13. https://doi.org/10.1088/1741-2560/3/1/R02

[11] Kang, H., Nam, Y., & Choi, S. (2009). Composite common spatial pattern for subject-to-subject transfer. *IEEE signal processing letters*, *16*(8), 683–686. https://doi.org/10.1109/LSP.2009.2022557

[12] Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., & Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE computational intelligence magazine*, *11*(1), 20–31. https://doi.org/10.1109/MCI.2015.2501545

[13] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, *15*(5), 56013. https://doi.org/10.1088/1741-2552/aace8c

[14] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., … & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, *38*(11), 5391–5420. https://doi.org/10.1002/hbm.23730

[15] Song, Y., Zheng, Q., Liu, B., & Gao, X. (2023). EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE transactions on neural systems and rehabilitation engineering*, *31*, 710–719. https://doi.org/10.1109/TNSRE.2022.3230250

[16] Barret, Z., Le Quoc, V. (2017). Neural architecture search with reinforcement learning. *International conference on learning representatoins* (Vol. 1, pp. 1–16). ICLR. https://arxiv.org/pdf/1611.01578

[17] Liu, H., Simonyan, K., & Yang, Y. (2018). *Darts: Differentiable architecture search*. https://doi.org/10.48550/arXiv.1806.09055

[18] Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized evolution for image classifier architecture search. *Proceedings of the AAAI conference on artificial intelligence* (pp. 4780–4789). PKP Publishing Services Network. https://doi.org/10.1609/aaai.v33i01.33014780

[19] Sun, J., Xie, J., & Zhou, H. (2021). EEG classification with transformer-based models. *2021 IEEE 3rd global conference on life sciences and technologies (Lifetech)* (pp. 92–93). IEEE. https://doi.org/10.1109/LifeTech52111.2021.9391844

[20] Zhu, Q., Zhao, X., Zhang, J., Gu, Y., Weng, C., & Hu, Y. (2023). *EEG2VEC: Self-supervised electroencephalographic representation learning*. https://doi.org/10.48550/arXiv.2305.13957

[21] Zhao, S., & Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 992–996). IEEE. 10.1109/ICASSP.2015.7178118

[22] Coretto, G. A. P., Gareis, I. E., & Rufiner, H. L. (2017). Open access database of eeg signals recorded during imagined speech. *12th international symposium on medical information processing and analysis* (Vol. 10160, p. 1016002). SPIE. https://doi.org/10.1117/12.2255697

[23] Torres-García, A. A., Reyes-García, C. A., Villaseñor-Pineda, L., & García-Aguilar, G. (2016). Implementing a fuzzy inference system in a multi-objective EEG channel selection model for imagined speech classification. *Expert systems with applications*, *59*, 1–12. https://doi.org/10.1016/j.eswa.2016.04.011

[24] Iqbal, S., Khan, Y. U., & Farooq, O. (2015). EEG based classification of imagined vowel sounds. *2015 2nd international conference on computing for sustainable global development (INDIACom)* (pp. 1591–1594). IEEE. https://arxiv.org/pdf/1904.05746

[25] Lee, S. H., Lee, M., Jeong, J. H., & Lee, S. W. (2019). Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery. *2019 IEEE international conference on systems, man and cybernetics (SMC)* (pp. 4409–4414). IEEE. https://doi.org/10.1109/SMC.2019.8914645

[26] Nieto, N., Peterson, V., Rufiner, H. L., Kamienkowski, J. E., & Spies, R. (2022). Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Scientific data*, *9*(1), 52. https://doi.org/10.1038/s41597-022-01147-2

[27] Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., … & Keutzer, K. (2019). FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search. *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10726–10734). IEEE. https://doi.org/10.1109/CVPR.2019.01099

[28] He, H., & Wu, D. (2020). Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE transactions on biomedical engineering*, *67*(2), 399–410. https://doi.org/10.1109/TBME.2019.2913914

[29] Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2013). Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, *112*, 172–178. https://doi.org/10.1016/j.neucom.2012.12.039

[30] Özdenizci, O., Wang, Y., Koike-Akino, T., & Erdoğmuş, D. (2020). Learning invariant representations from EEG via adversarial inference. *IEEE access*, *8*, 27074–27085. https://doi.org/10.1109/ACCESS.2020.2971600

[31] Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 97–105). Lille, France: PMLR. https://proceedings.mlr.press/v37/long15.html

[32] Zhao, H., Zheng, Q., Ma, K., Li, H., & Zheng, Y. (2020). Deep representation-based domain adaptation for nonstationary EEG classification. *IEEE transactions on neural networks and learning systems*, *32*(2), 535–545. https://doi.org/10.1109/TNNLS.2020.3010780

[33] Zanini, P., Congedo, M., Jutten, C., Said, S., & Berthoumieu, Y. (2018). Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE transactions on biomedical engineering*, *65*(5), 1107–1116. https://doi.org/10.1109/TBME.2017.2742541

[34] Rodrigues, P. L. C., Jutten, C., & Congedo, M. (2019). Riemannian procrustes analysis: Transfer learning for brain–computer interfaces. *IEEE transactions on biomedical engineering*, *66*(8), 2390–2401. https://doi.org/10.1109/TBME.2018.2889705

[35] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). MnasNet: Platform-aware neural architecture search for mobile. *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2815–2823). IEEE. https://doi.org/10.1109/CVPR.2019.00293

[36] Bender, G., Kindermans, P.-J., Zoph, B., Vasudevan, V., & Le, Q. (2018). Understanding and simplifying one-shot architecture search. *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 550–559). PMLR. https://proceedings.mlr.press/v80/bender18a/bender18a.pdf

[37] Guo, Z., Zhang, X., Mu, H., Heng, W., Liu, Z., Wei, Y., & Sun, J. (2020). Single path one-shot neural architecture search with uniform sampling. *Computer vision- ECCV 2020* (pp. 544–560). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58517-4_32

[38] Holland, J. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI. https://doi.org/10.7551/mitpress/1090.001.0001

[39] Eberhart, R., & Kennedy, J. (1995). Particle swarm optimization. *Proceedings of the IEEE international conference on neural networks* (Vol. 4, pp. 1942–1948). IEEE. https://doi.org/10.1109/ICNN.1995.488968

[40] Storn, R., & Price, K. (1997). Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

[41] Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

[42] Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE transactions on evolutionary computation*, *1*(1), 53–66. https://doi.org/10.1109/4235.585892

[43] Stanley, K. O., Clune, J., Lehman, J., & Miikkulainen, R. (2019). Designing neural networks through neuroevolution. *Nature machine intelligence*, *1*(1), 24–35. https://doi.org/10.1038/s42256-018-0006-z

[44] Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary computation*, *10*(2), 99–127. https://doi.org/10.1162/106365602320169811

[45] Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y. L., Tan, J., …& Kurakin, A. (2017). Large-scale evolution of image classifiers. *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 2902–2911). PMLR. https://proceedings.mlr.press/v70/real17a.html

[46] Aler, R., Galván, I. M., & Valls, J. M. (2012). Applying evolution strategies to preprocessing EEG signals for brain–computer interfaces. *Information sciences*, *215*, 53–66. https://doi.org/10.1016/j.ins.2012.05.012

[47] Zhang, R., Xu, P., Liu, T., Zhang, Y., Guo, L., Li, P., & Yao, D. (2013). Local temporal correlation common spatial patterns for single trial EEG classification during motor imagery. *Computational and mathematical methods in medicine*, *2013*(1), 591216. https://doi.org/10.1155/2013/591216

[48] Ilyas, M. Z., Saad, P., & Ahmad, M. I. (2015). A survey of analysis and classification of eeg signals for brain-computer interfaces. *2015 2nd international conference on biomedical engineering (ICOBE)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICoBE.2015.7235129

[49] Salmelin, R., & Hari, R. (1994). Spatiotemporal characteristics of sensorimotor neuromagnetic rhythms related to thumb movement. *Neuroscience*, *60*(2), 537–550. https://doi.org/10.1016/0306-4522(94)90263-1

[50] Crone, N. E., Hao, L., Hart, J., Boatman, D., Lesser, R. P., Irizarry, R., & Gordon, B. (2001). Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology*, *57*(11), 2045–2053. https://doi.org/10.1212/WNL.57.11.2045

[51] Bastiaansen, M., & Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. In *Event-related dynamics of brain oscillations* (Vol. 159, pp. 179–196). Elsevier. https://doi.org/10.1016/S0079-6123(06)59012-0

[52] Pfurtscheller, G., & Lopes Da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical neurophysiology*, *110*(11), 1842–1857. https://doi.org/10.1016/S1388-2457(99)00141-8

[53] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141). IEEE. https://doi.org/10.1109/CVPR.2018.00745

[54] Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 464–468). New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2074

[55] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

[56] Ramoser, H., Müller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE transactions on rehabilitation engineering*, *8*(4), 441–446. https://doi.org/10.1109/86.895946

[57] Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2008). Filter bank common spatial pattern (FBCSP) in brain-computer interface. *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 2390–2397). IEEE. https://doi.org/10.1109/IJCNN.2008.4634130

[58] Ingolfsson, T. M., Hersche, M., Wang, X., Kobayashi, N., Cavigelli, L., & Benini, L. (2020). EEG-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. *2020 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 2958–2965). IEEE. https://doi.org/10.1109/SMC42975.2020.9283028

[59] Kingma, D. P., & Ba, J. L. (2014). Adam: Amethod for stochastic optimization. *Proceedings of the 3rd international conference on learning representations (ICLR)* (pp. 1–15). ICLR. https://jjcurtin.quarto.pub/010_neural_networks-6260/pdfs/kingma_adam_optimizer.pdf

[60] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 3319–3328). PMLR. https://doi.org/10.5555/3305890.3306024

[61] Mersov, A. M., Jobst, C., Cheyne, D. O., & De Nil, L. (2016). Sensorimotor oscillations prior to speech onset reflect altered motor networks in adults who stutter. *Frontiers in human neuroscience*, *10*, 1–16. https://doi.org/10.3389/fnhum.2016.00443