

Article

Composite Gramian Angular Field for Time-Series Classification

Pero Bogunović¹, Saša Mladenović^{2,*}  and Andrina Granić^{2,*} 

¹ Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Ruđera Boškovića 32, 21000 Split, Croatia ; pero.bogunovic@fesb.hr

² Faculty of Science, University of Split, Ruđera Boškovića 33, 21000 Split, Croatia

* Correspondence: sasa.mladenovic@pmfst.hr (S.M.); andrina.granic@pmfst.hr (A.G.)

Abstract

Gramian Angular Field (GAF) encodings transform time series into two-dimensional images suitable for convolutional neural network (CNN) classification. Existing applications typically use either the Gramian Angular Summation Field (GASF) or the Gramian Angular Difference Field (GADF) independently, although these two encodings capture complementary pairwise angular relationships. This paper proposes the Composite Gramian Angular Field (CGAF), a single-image time-series representation obtained by a weighted algebraic combination of the summation and difference GAF components. The weights are optimised using coarse grid search followed by Gaussian-process Bayesian refinement, with all candidate evaluation restricted to training-only inner validation partitions. The selected weights are frozen before held-out test evaluation. CGAF produces a single encoded output image (approximately 0.08 MB, compared with approximately 0.16 MB for retaining separate GASF and GADF images) and encodes at 5.9 ± 0.3 ms per sample. We evaluate CGAF in three domain-specific settings—EEG cognitive engagement, PTB-DB heartbeat classification, and FordA automotive fault detection—and on a selected subset of 20 datasets from the UCR Time Series Classification Archive. The method is compared with GASF, GADF, recurrence plots, spectrogram-based encodings, and non-image time-series baselines including SVM, ResNet-1D, InceptionTime, and ROCKET. On the evaluated datasets, CGAF consistently improves over the individual GASF and GADF encodings. It achieves macro-F1 = 0.867 ± 0.027 on the EEG pilot study, heartbeat-segment-level macro-F1 = 0.941 ± 0.018 on PTB-DB, and test accuracy = 91.2% on FordA. Because patient identifiers are unavailable for PTB-DB, that result does not establish patient-level generalisation. On the selected UCR subset, CGAF outperforms both GASF and GADF on all 20 datasets. It achieves the best overall accuracy among all evaluated methods on 14 of 20 datasets, whereas ROCKET achieves the best overall accuracy on the remaining six datasets. The results suggest that algebraic integration of summation-based and difference-based angular dependencies can improve image-based time-series classification without modifying the CNN backbone or adding gradient-trained parameters. The EEG results should be interpreted as pilot evidence, whereas broader generalisation requires evaluation on the full UCR/UEA archive, additional biomedical cohorts, and further backbone architectures.



Academic Editors: Zoran H. Perić,
Tomas Skovranek, Vlado Delić,
Vladimir Despotovic and Stefan Panic

Received: 7 June 2026

Revised: 28 June 2026

Accepted: 29 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Gramian Angular Field; composite encoding; time-series image classification; weighted algebraic fusion; EEG; ECG; deep learning; signal processing; coarse grid search; Bayesian optimisation; UCR benchmark

1. Introduction

Time-series classification is a central task in many application domains, including biomedical signal analysis, industrial monitoring, mechanical fault diagnosis, and human–computer interaction. Signals such as electroencephalogram (EEG), electrocardiogram (ECG), vibration, and sensor measurements often exhibit non-stationarity, noise, variable recording lengths, and limited availability of labelled data in specialised domains [1,2]. These properties make direct classification challenging, particularly when the goal is to use deep learning models under realistic data constraints.

One practical approach is to transform one-dimensional time series into two-dimensional image representations and then apply convolutional neural networks (CNNs). Image-based encodings provide fixed-size inputs, expose pairwise or time-frequency relationships in a spatial form, and make it possible to use CNN architectures and transfer-learning strategies originally developed for computer vision. However, the effectiveness of this approach depends strongly on the chosen encoding. Different encodings emphasise different aspects of the underlying signal, and no single representation can be assumed to be optimal across domains.

The Gramian Angular Field (GAF) framework [3] is a widely used image-encoding method for time series. After normalisation and polar transformation, GAF represents temporal samples through pairwise angular relationships. Two standard variants are commonly used: the Gramian Angular Summation Field (GASF), based on $\cos(\phi_i + \phi_j)$, and the Gramian Angular Difference Field (GADF), based on $\sin(\phi_i - \phi_j)$. These two components capture different angular relationships: the summation field emphasises symmetric pairwise structure, whereas the difference field emphasises antisymmetric directional relationships. Existing applications often use GASF and GADF independently, retain them as separate channels, or combine representations at later stages [3–6]. These approaches differ from CGAF because they do not form a single algebraically integrated image in which the weighted summation-based and difference-based angular relationships are presented jointly to the convolutional filters.

This paper proposes the Composite Gramian Angular Field (CGAF), a single-image time-series representation obtained by a weighted algebraic combination of the GASF and GADF components. The objective is not to introduce a more complex CNN architecture, but to improve the information presented to a standard CNN backbone. By combining summation-based and difference-based angular dependencies in a single image, CGAF allows the same convolutional filters to operate on both complementary structures without requiring multi-channel stacking, late fusion, or additional gradient-trained parameters in the classifier.

1.1. Research Gap

Although GAF-based encodings are well established, the two standard GAF components are usually treated as alternative representations rather than as complementary components of a unified image encoding. GASF and GADF encode non-identical angular structures, but using only one of them may restrict the information available to the classifier. Late-fusion approaches can combine decisions from multiple encodings, but they require separate models and do not allow the CNN to learn local joint patterns formed by both components within the same input image.

To the best of our knowledge, limited prior work has investigated a weighted single-image integration of the summation and difference GAF components before CNN-based classification. This motivates the proposed CGAF representation and its evaluation across domain-specific and benchmark time-series datasets.

1.2. Contributions

The contributions of this paper are as follows:

1. Composite single-image GAF representation. CGAF defines an element-wise weighted algebraic combination of the summation and difference GAF components:

$$G_{\text{CGAF}}(i, j) = w_s \cos(\phi_i + \phi_j) + w_d \sin(\phi_i - \phi_j).$$

The representation is computed as a single image and therefore requires no modification of the CNN input architecture, no late-fusion ensemble, and no additional gradient-trained parameters in the classifier.

2. Data-driven optimisation of component weights. The weights w_s and w_d are optimised using coarse grid search followed by Gaussian-process Bayesian refinement. Candidate evaluation is restricted to training-only inner validation partitions, and the selected weights are frozen before held-out test evaluation.
3. Formal characterisation of complementary GAF components. We provide a mathematical characterisation showing that the summation and difference GAF components correspond to linearly independent trigonometric structures. This establishes a rationale for combining them, while avoiding the stronger and unnecessary claim that the composite representation must dominate either component for every dataset or classifier.
4. Controlled empirical evaluation. CGAF is evaluated in three domain-specific settings—EEG cognitive engagement, PTB-DB heartbeat classification, and FordA automotive fault detection—and on a selected subset of 20 datasets from the UCR Time Series Classification Archive. The comparison includes GASE, GADF, recurrence plots, spectrogram-based encodings, and non-image time-series baselines including SVM, ResNet-1D [7], InceptionTime [8], and ROCKET [9].
5. Reproducibility-oriented experimental protocol. The study reports preprocessing steps, fold-level validation procedures, weight-search details, computational cost, and implementation choices. The EEG dataset is treated as pilot evidence, the PTB-DB evaluation is explicitly limited to heartbeat-segment-level generalisation, and the public FordA and selected UCR datasets provide broader evaluation under the reported experimental protocols.

The remainder of the paper is organised as follows. Section 2 reviews related work on time-series image encoding and GAF-based classification. Section 3 presents the proposed CGAF method. Section 4 describes the experimental setup. Section 5 reports the results. Section 6 discusses the findings and limitations. Section 7 concludes the paper.

2. Related Work

2.1. Time-Series Image Encoding Methods

Transforming one-dimensional time series into two-dimensional representations has become a common strategy for applying convolutional neural networks (CNNs) to signal classification tasks. Several families of encodings have been used for this purpose. Recurrence plots (RP) visualise repeated states and phase-space recurrences in dynamical systems [10] and have been applied to EEG and activity-recognition problems [11]. Markov Transition Fields (MTF) represent transition probabilities between quantised amplitude states while preserving temporal ordering [3]. Spectrograms and wavelet-based representations provide time-frequency views of signals and are particularly useful when spectral content is discriminative [12]. Symbolic Aggregate approXimation (SAX) and Bag-of-Words representations provide interpretable symbolic encodings, although they may lose fine-grained temporal variation due to discretisation [13].

Among image-based encodings, Gramian Angular Fields (GAF) are particularly relevant because they encode pairwise angular relationships between temporal samples after normalisation and polar transformation [3]. This makes GAF representations suitable for CNNs while preserving an explicit relationship between temporal ordering and image structure. GAF-based methods have been used in biomedical, industrial, and financial time-series applications, including ECG analysis, EEG classification, fault detection, and financial forecasting [4–6,14].

Image-based approaches should also be considered in relation to strong one-dimensional time-series classifiers. Residual networks for time series [7], InceptionTime [8], and ROCKET [9] have shown strong performance on benchmark time-series classification problems without requiring image conversion. These methods are therefore used in this paper as non-image reference baselines rather than as direct competitors within the same representation family. The purpose of CGAF is not to replace one-dimensional classifiers, but to improve the information content of GAF-style image representations when a CNN-based image-classification pipeline is used.

2.2. GAF-Based Signal Classification

The original GAF framework defines two related but distinct image encodings: the Gramian Angular Summation Field (GASF), based on $\cos(\phi_i + \phi_j)$, and the Gramian Angular Difference Field (GADF), based on $\sin(\phi_i - \phi_j)$ [3]. GASF produces a symmetric matrix and emphasises summation-based pairwise angular structure, whereas GADF produces an antisymmetric matrix and emphasises directional angular differences. Both representations are derived from the same polar embedding of the time series, but they expose different pairwise relationships to the classifier.

Prior studies have typically selected one of these encodings for a specific task. For example, GAF-based representations have been used for ECG classification [4,15], EEG-related classification tasks [5,6], and other time-series domains. Other studies combine multiple time-series image encodings, such as GAF and MTF, in multi-channel or task-specific image-classification pipelines [16]. However, such approaches often keep the representations as separate channels or combine them at later stages, rather than forming a single algebraically integrated image. In such cases, the convolutional filters do not operate on a single image in which summation-based and difference-based angular relationships are jointly represented.

Prior work has most commonly used GASF and GADF as separate alternatives, as multi-channel inputs, or as components in ensemble settings. In contrast, this study focuses on a weighted single-image integration of the two GAF components before CNN-based classification. This motivates the proposed CGAF representation and its evaluation across domain-specific and benchmark time-series datasets.

2.3. Modern Deep Time-Series Representation Learning

Modern deep time-series representation learning includes transformer architectures, self-supervised methods [17], and foundation models. PatchTST [18], TimesNet [19], and MOMENT [20] established influential patch-based, temporal-variation, and pretrained modelling approaches. More recent classification-oriented developments include Mantis, a contrastively pretrained Vision Transformer-based foundation model [21]; TimesBERT, which uses BERT-style masked and functional-token prediction for time-series understanding tasks [22]; TiCT, which performs in-context classification after synthetic pretraining [23]; and UniShape, which introduces shape-aware adapters and prototype-based pretraining for transferable time-series classification [24]. These models demonstrate the rapid development of pretrained and in-context approaches for classification. They were not exper-

imentally evaluated in the present study, and their inclusion here provides an updated account of the field rather than evidence of comparative CGAF performance.

The present work addresses a narrower representation-level question: whether algebraic integration of GASF and GADF improves a CNN-based GAF image-classification pipeline. PatchTST, TimesNet, and MOMENT were not included in the experimental evaluation. Consequently, the reported results do not establish the performance of CGAF relative to transformer-based or foundation-model time-series classifiers. A fair comparison would require harmonised data splits, preprocessing, classification heads, hyperparameter budgets, and, where applicable, pretraining and fine-tuning conditions. Such an evaluation is reserved for future work.

2.4. Positioning of CGAF

The proposed method is motivated by the observation that GASF and GADF encode non-identical angular structures derived from the same time series. From a multi-view learning perspective, combining complementary views can improve classification when the views provide compatible and non-redundant information [25]. In the GAF setting, the summation and difference components can be interpreted as two complementary angular views. Section 3.4 formalises this relationship by showing that the two components are linearly independent trigonometric structures.

Table 1 summarises the position of CGAF relative to common GAF-based alternatives. The key distinction is that CGAF performs a weighted algebraic integration of the GASF and GADF components into a single image before CNN training, with weights optimised using training data.

Table 1. Conceptual positioning of CGAF relative to common GAF-based representation strategies. “Algebraic integration” denotes mathematical integration of components before CNN training. “Optimised weights” denotes whether the relative contribution of components is selected using training data. “Single image” denotes whether the method produces one two-dimensional matrix as CNN input. The table is intended as a methodological comparison, not as an exhaustive taxonomy of all GAF-based approaches.

Method	Fusion Type	CNN Input	Algebraic Integration	Optimised Weights
GASF only [3]	None	Single image	No	No
GADF only [3]	None	Single image	No	No
Late fusion/ensemble	Decision-level	Separate images	No	No
GASF + GADF stack	Pre-CNN stacking	Multi-channel image	No	No
Fixed-weight composite	Pre-CNN fusion	Single image	Yes	No
CGAF (ours)	Pre-CNN fusion	Single image	Yes	Yes

3. Proposed Method: CGAF

3.1. Preprocessing and Polar Embedding

Let $x = (x_1, \dots, x_N) \in \mathbb{R}^N$ denote a univariate time series after dataset-specific artefact removal, where applicable (Section 4.2). Each sample is normalised independently using min–max normalisation:

$$\tilde{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \quad \tilde{x}_i \in [0, 1], \quad (1)$$

where $x_{\min} = \min_k x_k$ and $x_{\max} = \max_k x_k$ for the given sample. The normalised values are then mapped to the interval $[-1, 1]$:

$$\hat{x}_i = 2\tilde{x}_i - 1, \quad \hat{x}_i \in [-1, 1]. \quad (2)$$

The polar embedding is defined as

$$\phi_i = \arccos(\hat{x}_i), \quad \phi_i \in [0, \pi]. \quad (3)$$

To obtain fixed-size images, Piecewise Aggregate Approximation (PAA) [26] is used to resample or compress each time series to a target length $m = 100$, unless otherwise stated in the dataset-specific protocol. The value $m = 100$ was selected as a fixed, dataset-independent compromise between temporal resolution and the quadratic computational cost of constructing an $m \times m$ GAF image. It was applied consistently across the compared image encodings to standardise their input dimensions. However, this choice is not claimed to be universally or dataset-specifically optimal, and different encodings may respond differently to changes in m . In the experiments reported here, normalisation is applied before PAA. For datasets with series shorter than the target length, padding is applied consistently within the same training and test protocol. When padding is performed using the minimum normalised value, the padded entries correspond to $\hat{x} = -1$ and therefore to the polar angle $\phi = \pi$. The padding convention used in each experiment is reported in the preprocessing details to avoid ambiguity. Additional preprocessing details are provided in Appendix A.

3.2. CGAF Definition

The standard Gramian Angular Field framework defines two related encodings: the Gramian Angular Summation Field (GASF),

$$G_{\text{GASF}}(i, j) = \cos(\phi_i + \phi_j), \quad (4)$$

and the Gramian Angular Difference Field (GADF),

$$G_{\text{GADF}}(i, j) = \sin(\phi_i - \phi_j). \quad (5)$$

GASF is symmetric, whereas GADF is antisymmetric. These two components encode different pairwise angular relationships derived from the same polar representation of the time series.

The proposed Composite Gramian Angular Field (CGAF) combines these two components into a single image:

$$G_{\text{CGAF}}(i, j) = w_s \cos(\phi_i + \phi_j) + w_d \sin(\phi_i - \phi_j), \quad w_s, w_d \in [0, 2]. \quad (6)$$

Here, w_s and w_d control the relative contribution of the summation and difference components, respectively. The weights are selected using training data only, as described in Section 3.5. The resulting CGAF image has the same spatial dimension as a single GAF image and therefore requires no change to the CNN input architecture.

The range of an unnormalised CGAF entry is bounded by

$$G_{\text{CGAF}}(i, j) \in [-(w_s + w_d), w_s + w_d]. \quad (7)$$

In the reported experiments, all encoded images are clipped to a fixed range and linearly rescaled to $[0, 255]$ before CNN input. The clipping range and the observed fraction of clipped pixels are reported together with the optimised weights. This makes the image conversion reproducible and avoids implicit dataset-dependent contrast scaling.

3.3. Illustrative Visual Comparison

Figure 1 illustrates the GASF, GADF, and CGAF encodings of the same representative EEG segment. The purpose of the figure is to show the qualitative difference between the three encodings, not to provide evidence of class separability by itself. GASF produces a symmetric image, GADF produces an antisymmetric image, and CGAF combines both angular components within a single image using the selected weights.

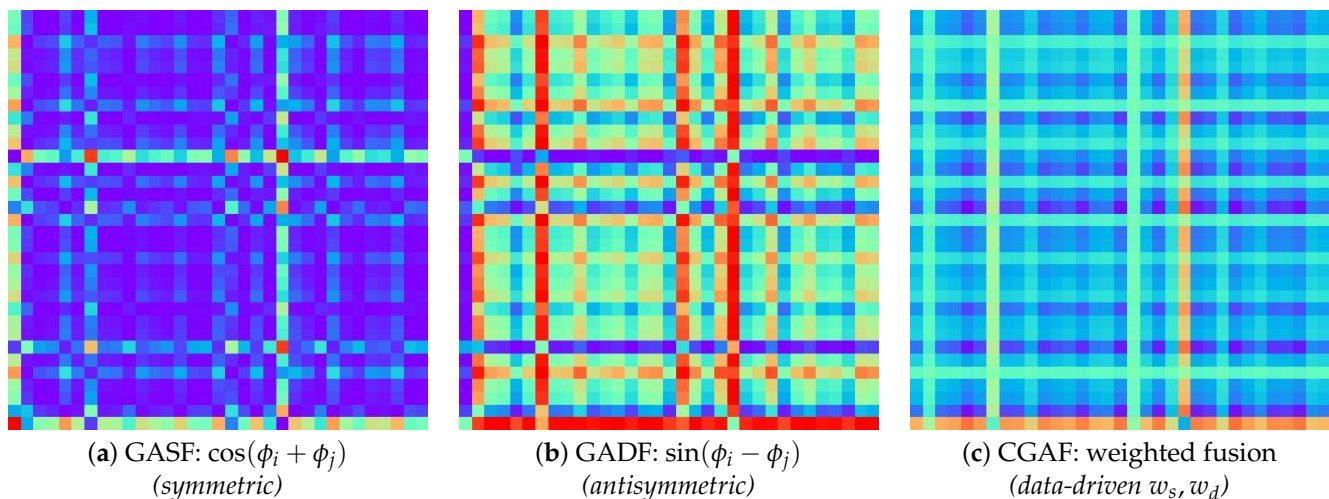


Figure 1. Illustrative comparison of GASF, GADF, and CGAF encodings for the same representative EEG segment. GASF produces a symmetric summation-based angular image, GADF produces an antisymmetric difference-based angular image, and CGAF combines both components into a single weighted image. The figure is intended to illustrate structural differences between the encodings; class separability is evaluated quantitatively in Section 5.

3.4. Theoretical Justification

The purpose of this subsection is limited to establishing the algebraic non-redundancy of the two GAF components. Their linear independence shows that the summation and difference functions represent distinct angular structures, but it does not establish that their composite contains more discriminative information, preserves additional signal characteristics, increases class separability, or improves classification performance. The practical value of their combination must therefore be determined empirically.

Using standard trigonometric identities, with $c_k = \cos \phi_k$ and $s_k = \sin \phi_k$, we have

$$\cos(\phi_i + \phi_j) = c_i c_j - s_i s_j, \quad (8)$$

$$\sin(\phi_i - \phi_j) = s_i c_j - c_i s_j. \quad (9)$$

These expressions correspond to different bilinear trigonometric structures over the same angular variables.

Proposition 1 (Linear independence of GAF angular components). *Let*

$$g_s(\phi_i, \phi_j) = \cos(\phi_i + \phi_j) \quad \text{and} \quad g_d(\phi_i, \phi_j) = \sin(\phi_i - \phi_j)$$

denote the summation and difference GAF components. Then, g_s and g_d are linearly independent as functions of (ϕ_i, ϕ_j) on $[0, \pi]^2$. Consequently, for any $w_s, w_d > 0$, the CGAF function

$$G_{\text{CGAF}}(\phi_i, \phi_j) = w_s g_s(\phi_i, \phi_j) + w_d g_d(\phi_i, \phi_j)$$

contains non-zero contributions from both linearly independent angular components.

Proof. Assume that there exist constants $\alpha, \beta \in \mathbb{R}$ such that

$$\alpha g_s(\phi_i, \phi_j) + \beta g_d(\phi_i, \phi_j) = 0$$

for all $(\phi_i, \phi_j) \in [0, \pi]^2$. Setting $(\phi_i, \phi_j) = (0, 0)$ gives

$$\alpha g_s(0, 0) + \beta g_d(0, 0) = \alpha = 0,$$

because $g_s(0, 0) = 1$ and $g_d(0, 0) = 0$. Setting $(\phi_i, \phi_j) = (\pi/2, 0)$ then gives

$$\alpha g_s(\pi/2, 0) + \beta g_d(\pi/2, 0) = \beta = 0,$$

because $g_s(\pi/2, 0) = 0$ and $g_d(\pi/2, 0) = 1$. Therefore, the only linear combination that vanishes identically is the trivial one, and g_s and g_d are linearly independent. $\square$

Remark 1 (Scope of the structural argument). *Proposition 1 establishes only that GASF and GADF are algebraically non-redundant functions of the angular variables. It does not establish orthogonality, information-theoretic independence, preservation of additional signal characteristics, greater mutual information with the class labels, or universally improved classification. Accordingly, the proposition is used only as structural motivation for combining the components. Evidence for the usefulness of CGAF is empirical and is evaluated in Sections 5 and 5.8.*

Structural Properties of CGAF

The CGAF representation inherits structural properties from both GAF components. First, for $w_d > 0$, CGAF is generally asymmetric because

$$\sin(\phi_i - \phi_j) = -\sin(\phi_j - \phi_i).$$

This means that the entries $G_{\text{CGAF}}(i, j)$ and $G_{\text{CGAF}}(j, i)$ may differ, allowing the image to encode ordered pairwise angular relationships.

Second, the diagonal of CGAF is determined only by the GASF component:

$$G_{\text{CGAF}}(i, i) = w_s \cos(2\phi_i) = w_s(2\hat{x}_i^2 - 1), \quad (10)$$

because $\sin(\phi_i - \phi_i) = 0$ and $\cos(\phi_i) = \hat{x}_i$ by definition of the polar embedding. Thus, the diagonal retains the GASF self-relation, whereas the GADF contribution affects only off-diagonal entries.

Third, the raw CGAF values are bounded by

$$G_{\text{CGAF}}(i, j) \in [-(w_s + w_d), w_s + w_d], \quad (11)$$

since both $\cos(\phi_i + \phi_j)$ and $\sin(\phi_i - \phi_j)$ lie in $[-1, 1]$. In the experiments, raw CGAF values are clipped to a fixed range and then linearly normalised before CNN input, as described in the experimental protocol. The observed clipping fractions are reported in Table 2.

Table 2. Derived weights (w_s, w_d) per dataset after full optimisation (magnitude-ratio initialisation, coarse grid search, and Bayesian refinement). $\bar{M}_s$ and $\bar{M}_d$ denote the mean absolute entry magnitudes of the GASF and GADF matrices over training samples. Clipped % denotes the fraction of pixels clipped at the fixed $[-2, 2]$ boundary before normalisation.

Dataset	$\bar{M}_s$ (Mean $\pm$ Std)	$\bar{M}_d$ (Mean $\pm$ Std)	w_s (Mean $\pm$ Std)	w_d (Mean $\pm$ Std)	Clipped %
EEG	0.431 $\pm$ 0.003	0.178 $\pm$ 0.002	0.31 $\pm$ 0.008	1.52 $\pm$ 0.009	0.3%
PTB-DB	0.401 $\pm$ 0.004	0.213 $\pm$ 0.003	0.45 $\pm$ 0.007	1.21 $\pm$ 0.008	0.1%
Ford A	0.389 $\pm$ 0.004	0.241 $\pm$ 0.003	0.57 $\pm$ 0.009	1.08 $\pm$ 0.010	0.2%

3.5. Weight Optimisation: Coarse Grid Search and Bayesian Refinement

The CGAF representation depends on two scalar weights, w_s and w_d , which control the relative contribution of the summation and difference GAF components. These weights are not gradient-trained parameters of the CNN classifier. Instead, they are selected before final CNN training using only the training data available within each validation fold.

3.5.1. Magnitude-Ratio Initialisation

Before grid search, a data-driven reference point ($\bar{w}_s, \bar{w}_d$) is computed from the training partition. Let $\bar{M}_s$ and $\bar{M}_d$ denote the mean absolute entry values of the GASF and GADF matrices over all training samples:

$$\bar{M}_s = \frac{1}{N_{\text{tr}} m^2} \sum_{n=1}^{N_{\text{tr}}} \sum_{i,j} |\cos(\phi_i^{(n)} + \phi_j^{(n)})|, \quad \bar{M}_d = \frac{1}{N_{\text{tr}} m^2} \sum_{n=1}^{N_{\text{tr}}} \sum_{i,j} |\sin(\phi_i^{(n)} - \phi_j^{(n)})|. \quad (12)$$

The corresponding reference weights are

$$\bar{w}_s = \frac{\bar{M}_d}{\bar{M}_s + \bar{M}_d}, \quad \bar{w}_d = \frac{\bar{M}_s}{\bar{M}_s + \bar{M}_d}. \quad (13)$$

This yields a normalised pair satisfying $\bar{w}_s + \bar{w}_d = 1$ and gives larger relative weight to the component with smaller mean absolute magnitude. The pair is used only as a descriptive reference point. The subsequent grid search is not restricted to its neighbourhood.

3.5.2. Search Grid

The primary search is performed over the bounded domain

$$\mathcal{W} = \{0.1, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0\} \times \{0.1, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0\},$$

yielding $7 \times 7 = 49$ candidate weight pairs. For each candidate pair, CGAF images are generated from the training partition, and the CNN is fine-tuned for a reduced five-epoch proxy run. The candidate achieving the highest validation macro-F1 is retained as the coarse-grid optimum. The reduced-epoch proxy is used only for weight selection; final model training uses the full training protocol described in Section 4.

Weight optimisation is performed exclusively within the training portion of each fold. The held-out test data are not used during magnitude-ratio calculation, grid search, Bayesian refinement, or validation-based weight selection.

The Bayesian-refinement stage uses the same training-only validation protocol as the preceding coarse-grid search. For EEG, a fixed subject-level inner training-validation split is constructed from the subjects in each outer training fold. For PTB-DB, a fixed stratified 80/20 training-validation split is constructed from the heartbeat segments in each outer training fold. For FordA and the UCR datasets, both optimisation stages use the same three-fold stratified cross-validation partitions within the official training partition, with

mean validation macro-F1 as the optimisation objective. After Bayesian refinement, the selected weight pair is frozen before evaluation on the corresponding outer test fold or official held-out test partition. No test observations are used during weight selection.

3.5.3. Bayesian Refinement

Starting from the best coarse-grid configuration, Gaussian-process Bayesian optimisation [27] is applied over a local neighbourhood

$$[w_s^* - 0.15, w_s^* + 0.15] \times [w_d^* - 0.15, w_d^* + 0.15],$$

clipped to the feasible range $[0, 2] \times [0, 2]$. The acquisition function is Expected Improvement (EI), and 20 additional evaluations are performed. This two-stage procedure provides a practical compromise between exhaustive search and manual weight selection.

The preliminary grid-density comparison is reported in Table 3.

Table 3. Grid density sensitivity. The 7×7 coarse grid achieved performance close to the denser 21×21 grid in preliminary experiments while requiring substantially fewer CNN proxy-training runs. Wall-clock times include fixed overhead for data loading and fold setup.

Grid	Candidates	EEG F1	PTB-DB F1	Wall-Clock (min/Fold)
21×21 ($\Delta = 0.1$)	441	0.867	0.941	~ 310
7×7 coarse grid	49	0.858	0.939	~ 37

3.5.4. Mapping Between λ and (w_s, w_d)

For interpretability in the ablation analysis, Section 5.8 also considers a constrained one-parameter version:

$$G_\lambda(i, j) = (1 - \lambda) \cos(\phi_i + \phi_j) + \lambda \sin(\phi_i - \phi_j), \quad \lambda \in [0, 1]. \quad (14)$$

This corresponds to setting $w_s = 1 - \lambda$ and $w_d = \lambda$, i.e., a normalised convex combination satisfying $w_s + w_d = 1$. This constrained model is a subset of the full CGAF search space. The full optimisation over $[0, 2] \times [0, 2]$ additionally allows amplification regimes in which $w_s + w_d > 1$.

In the reported experiments, the best constrained value was near $\lambda \approx 0.71$, while the unconstrained Bayesian refinement selected weights such as $w_s = 0.31$ and $w_d = 1.52$ for the EEG dataset. This indicates that the validation procedure selected a solution outside the convex-combination subspace. The benefit of this additional flexibility is evaluated empirically in the ablation experiments rather than assumed theoretically.

3.5.5. Empirical Optimal Weights

Across the three domain-specific datasets, the optimised difference-component weight w_d was larger than the summation-component weight w_s . This pattern is reported descriptively and should not be interpreted as a universal property of CGAF. Table 2 reports the final optimised fold-level weights and clipping fractions.

3.6. Complete Algorithm

The fold-level CGAF encoding and evaluation procedure is formalised in Algorithm 1.

Remark 2 (Clip bound). After CGAF construction, raw values are clipped to $[-2, 2]$ and linearly normalised to $[0, 255]$ before CNN input. This fixed clipping range was chosen because all optimised weight configurations in the reported domain-specific experiments satisfied $w_s + w_d < 2$. The

observed clipped-pixel fractions were negligible, as reported in Table 2. For manually set weights or future datasets with $w_s + w_d > 2$, the clipping range should be adjusted to $[-(w_s + w_d), w_s + w_d]$ or otherwise reported explicitly.

Complexity.

For a target length m , constructing a CGAF image requires $\mathcal{O}(m^2)$ operations per sample. The measured encoding time was 5.9 ± 0.3 ms per sample on CPU only using an Intel i7-9700K processor, single-threaded execution, $m = 100$, and batch size 1. Weight optimisation adds a one-time per-fold cost because each candidate configuration requires a proxy CNN training run. This cost is separated from inference-time encoding and is analysed in Section 5.9.

Algorithm 1 CGAF encoding and fold-level evaluation.

Require: Training partition $\mathcal{D}_{\text{tr}}$, validation partition $\mathcal{D}_{\text{val}}$, held-out test partition $\mathcal{D}_{\text{te}}$, target length m

Ensure: Fold-specific weights (w_s, w_d) and test performance estimate

- 1: Preprocess and normalise samples within the fold according to Equations (1) and (2).
 - 2: Apply PAA to obtain length m and compute polar angles $\phi_i = \arccos(\hat{x}_i)$.
 - 3: **for** each candidate (w_s, w_d) in the coarse grid **do**
 - 4: Generate CGAF images for $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{val}}$ using Equation (6).
 - 5: Fine-tune the CNN using the reduced-epoch proxy protocol on $\mathcal{D}_{\text{tr}}$.
 - 6: Evaluate validation macro-F1 on $\mathcal{D}_{\text{val}}$.
 - 7: **end for**
 - 8: Select the best coarse-grid weight pair according to validation macro-F1.
 - 9: Refine the selected pair using Gaussian-process Bayesian optimisation with Expected Improvement, using the same training-only validation split(s) and macro-F1 objective as in the coarse-grid stage.
 - 10: Freeze the final fold-specific weights (w_s, w_d) .
 - 11: Generate CGAF images for the full fold training partition using the frozen weights.
 - 12: Train the final CNN using the full training protocol.
 - 13: Generate CGAF images for the held-out test partition using the same frozen weights.
 - 14: Evaluate the final trained model once on the held-out test partition.
 - 15: **return** (w_s, w_d) and the fold-level test performance.
-

The full preprocessing, encoding, and classification workflow is summarised in Figure 2.

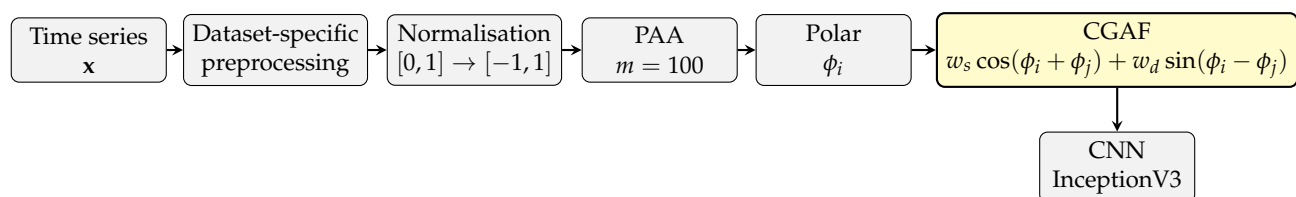


Figure 2. CGAF processing pipeline. After dataset-specific preprocessing, where applicable, each time series is normalised before PAA, transformed into polar angles, and encoded as a weighted algebraic combination of the summation and difference GAF components. The weights (w_s, w_d) are selected exclusively from the available training data using the corresponding inner validation protocol and are then frozen before evaluation on held-out test data.

4. Experimental Setup

4.1. Datasets

The proposed CGAF representation was evaluated on three domain-specific datasets and on a selected subset of 20 datasets from the UCR Time Series Classification Archive. The domain-specific datasets were used to assess CGAF in biomedical, educational, and

industrial signal settings, whereas the UCR subset was used to provide broader benchmark coverage across different time-series domains, lengths, and class counts.

4.1.1. Dataset 1: EEG Cognitive Engagement (Pilot Study; Proprietary, Ethics-Approved)

The EEG dataset consists of 500 anonymised EEG segments derived from school-based science-popularisation activities involving 100 primary-school pupils aged 8–11 years (52 female and 48 male). The activities included arithmetic and logical problem-solving tasks and used the NeuroSky MindWave device, a consumer-grade brain–computer interface device intended for educational and demonstration purposes rather than medical diagnosis. Signals were recorded using the NeuroSky MindWave Mobile device with a dry FP1 electrode, 512 Hz sampling frequency, and 12-bit resolution. Each segment has a duration of 3 s, corresponding to 1536 samples.

The classification task is binary: Engaged versus Distracted. Labels were assigned by a trained educational psychologist based on direct behavioural observation and task-performance scoring. The labels were assigned independently of the physiological features used by the classifier, reducing the risk of circularity between signal-derived features and target labels. The dataset contains 262 engaged and 238 distracted segments.

This dataset is treated as a pilot study. Although subject-independent cross-validation is used, the cohort contains only 100 participants, and each test fold contains a limited number of samples. Therefore, the EEG results are reported as preliminary evidence for the feasibility of CGAF in this setting and should not be interpreted as population-level generalisation. Broader cross-domain evidence is provided primarily by the public FordA and selected UCR experiments, while PTB-DB provides an exploratory heartbeat-segment-level biomedical evaluation.

Written parental consent was obtained, and the data were pseudonymised before analysis.

4.1.2. Dataset 2: Exploratory PTB-DB Heartbeat Data

The exploratory PTB analysis used 1000 heartbeat samples obtained from the ECG Heartbeat Categorization Dataset available through Kaggle. The files `ptbdb_abnormal.csv` and `ptbdb_normal.csv` contain preprocessed and segmented Lead-II heartbeat signals derived from the PTB Diagnostic ECG Database [28,29]. The analysed subset comprised 543 abnormal and 457 normal heartbeat samples. Each row represents one fixed-length heartbeat sampled at 125 Hz, with the class label stored in the final column. No ECG lead was selected by the authors. Because the distributed CSV files do not contain patient identifiers, patient-independent partitioning could not be verified. The PTB results are therefore interpreted at the heartbeat-segment level and not as evidence of patient-level generalisation.

The dataset was included as a public biomedical benchmark. The VGG16 result reported by Elmir et al. [4] is included in Table 4 only as a contextual reference because the experimental protocols are not identical.

Table 4. Heartbeat-segment-level classification performance on the PTB Diagnostic ECG dataset (binary: myocardial infarction vs. healthy). Results are reported as mean $\pm$ standard deviation over stratified 10-fold cross-validation repeated three times. Because patient identifiers were unavailable, these results do not represent patient-independent evaluation. Best values among InceptionV3-based methods are shown in bold. Rows marked with ‡ are contextual references from a different protocol and are not used for direct statistical comparison.

Method	Backbone	Accuracy	Precision	Recall	F1	$p_{w/d}$
SVM	—	0.716 $\pm$ 0.051	0.722 $\pm$ 0.049	0.713 $\pm$ 0.052	0.709 $\pm$ 0.052	<0.001/1.9
1D-CNN	—	0.782 $\pm$ 0.041	0.788 $\pm$ 0.039	0.779 $\pm$ 0.042	0.775 $\pm$ 0.042	<0.001/1.6
ResNet-1D	—	0.811 $\pm$ 0.037	0.817 $\pm$ 0.035	0.808 $\pm$ 0.038	0.804 $\pm$ 0.037	<0.001/1.4
InceptionTime	—	0.827 $\pm$ 0.034	0.833 $\pm$ 0.032	0.824 $\pm$ 0.035	0.819 $\pm$ 0.034	0.001/1.2
ROCKET	—	0.836 $\pm$ 0.031	0.841 $\pm$ 0.030	0.832 $\pm$ 0.032	0.828 $\pm$ 0.031	0.001/1.1
RP	InceptionV3	0.857 $\pm$ 0.037	0.863 $\pm$ 0.035	0.854 $\pm$ 0.038	0.851 $\pm$ 0.037	0.002/0.9
MTF	InceptionV3	0.871 $\pm$ 0.030	0.877 $\pm$ 0.028	0.868 $\pm$ 0.031	0.863 $\pm$ 0.030	0.008/0.8
SAX+BoW	InceptionV3	0.631 $\pm$ 0.048	0.638 $\pm$ 0.046	0.628 $\pm$ 0.049	0.623 $\pm$ 0.048	<0.001/2.8
Spectrogram	InceptionV3	0.748 $\pm$ 0.042	0.754 $\pm$ 0.041	0.745 $\pm$ 0.043	0.741 $\pm$ 0.042	<0.001/1.8
GASF	InceptionV3	0.889 $\pm$ 0.026	0.894 $\pm$ 0.024	0.885 $\pm$ 0.027	0.881 $\pm$ 0.026	0.015/0.7
GADF	InceptionV3	0.901 $\pm$ 0.023	0.906 $\pm$ 0.021	0.897 $\pm$ 0.024	0.894 $\pm$ 0.023	0.031/0.6
Contextual reference; different protocol, direct statistical comparison not valid						
Elmir et al. [4] ‡	VGG16	0.8917	—	—	0.8898	ref.
CGAF (ours)	VGG16	0.936 $\pm$ 0.019	0.941 $\pm$ 0.017	0.932 $\pm$ 0.020	0.930 $\pm$ 0.019	—
CGAF (ours)	InceptionV3	0.948 $\pm$ 0.018	0.953 $\pm$ 0.016	0.944 $\pm$ 0.019	0.941 $\pm$ 0.018	—

4.1.3. Dataset 3: FordA Fault Detection (Public, Binary Classification)

The FordA Fault Detection dataset was obtained from the UCR Time Series Classification Archive [30]. It contains 3601 training samples and 1320 test samples for binary classification. Each time series consists of 500 measurements of engine sensor noise. In this study, the standard UCR training–test split was retained.

FordA appears in two parts of the paper. First, it is included as a domain-specific industrial fault-detection dataset, where full per-method results are reported. Second, it appears as one of the 20 UCR benchmark datasets, where it supports comparison under the unified UCR protocol. These two uses refer to the same final held-out test partition and are therefore complementary rather than independent experiments.

4.1.4. Datasets 4–23: UCR Time Series Classification Archive

To evaluate broader benchmark behaviour, CGAF was additionally evaluated on 20 datasets from the UCR Time Series Classification Archive [30]. The selected datasets cover several domains, including ECG, energy, motion, industrial, plant, seismic, household, medical, and synthetic time series. They also vary in total number of samples, series length, and number of classes. This selection was intended to provide heterogeneous benchmark coverage and to reduce dependence on a single application domain. Accordingly, the UCR subset was used as an exploratory cross-domain validation set rather than as a comprehensive benchmark over the full UCR/UEA archive.

Table 5 lists the selected UCR datasets and their main properties.

Table 5. Selected UCR datasets used for broader evaluation. *N*: total number of samples; *L*: series length; *C*: number of classes.

#	Dataset	<i>N</i>	<i>L</i>	<i>C</i>	Domain
1	ECG5000	5000	140	5	ECG
2	ElectricDevices	16,637	96	7	Energy
3	FaceAll	2250	131	14	Image-derived
4	GunPoint	200	150	2	Motion
5	ItalyPowerDemand	1096	24	2	Energy
6	MoteStrain	1272	84	2	Sensor
7	NonInvasiveFetalECG_Thorax1	3765	750	42	ECG
8	OSULeaf	442	427	6	Plant
9	SonyAIBORobotSurface1	621	70	2	Robot
10	SwedishLeaf	1125	128	15	Plant
11	TwoPatterns	5000	128	4	Synthetic
12	UWaveGestureLibraryAll	4478	945	8	Gesture
13	Wafer	7164	152	2	Industrial
14	FordA	4921	500	2	Automotive
15	Earthquakes	461	512	2	Seismic
16	FreezerRegularTrain	3000	301	2	Household
17	HandOutlines	1370	2709	2	Shape
18	MiddlePhalanxOutlineAgeGroup	554	80	3	Medical
19	SmoothSubspace	300	15	3	Synthetic
20	Yoga	3300	426	2	Motion

4.2. Artefact Removal

4.2.1. EEG Artefact Removal

Because only one EEG channel was available, conventional multichannel ICA and VEOG-based ocular correction were not used. Each 3 s EEG segment was filtered using a zero-phase fifth-order Butterworth band-pass filter from 0.5 to 50 Hz. A segment was rejected in its entirety when the maximum absolute amplitude of the filtered signal exceeded 100 μ V. This segment-level amplitude gate was intended to exclude large-amplitude transient contamination, including prominent frontal blink artefacts; lower-amplitude ocular contamination may nevertheless remain. Given the consumer-grade single-channel acquisition, this procedure should be interpreted as pragmatic artefact rejection rather than clinical-grade source separation.

4.2.2. PTB-DB Preprocessing

No raw-signal filtering, QRS detection, artefact rejection, lead selection, or heartbeat segmentation was performed by the authors. The PTB-DB samples were obtained as pre-processed, fixed-length heartbeat segments from the Kaggle ECG Heartbeat Categorization Dataset. The study-specific transformations comprised per-sample min–max normalisation, mapping to $[-1, 1]$, PAA resampling to $m = 100$, and subsequent GAF or CGAF encoding.

4.2.3. UCR Datasets

No additional artefact removal was applied to the UCR datasets. The datasets were used as provided in the UCR archive, except for the fixed-length resampling and image-encoding steps described in the preprocessing protocol.

4.3. Training/Test Splits

For the EEG dataset, subject-independent stratified 10-fold cross-validation with three repeats was used. Subject independence was enforced so that segments from the same participant did not appear simultaneously in training and test partitions.

For PTB-DB, heartbeat-segment-level stratified 10-fold cross-validation repeated three times was used. Within each outer training fold, a fixed stratified 80/20 training–validation split was used for both coarse-grid weight selection and Bayesian refinement. The selected weight pair was then frozen before evaluation on the corresponding outer test fold. Because the distributed CSV files did not contain patient identifiers, patient-independent partitioning could not be enforced; consequently, the results do not establish patient-level generalisation.

For FordA and the UCR datasets, the standard UCR training–test split was retained. Within the training partition, stratified cross-validation was used for weight optimisation and hyperparameter selection. Final performance was evaluated only on the standard held-out test partition.

4.4. Baselines

The proposed method was compared with non-image time-series classifiers, alternative image encodings, and the individual GAF components.

1. SVM (RBF): A 22-dimensional handcrafted feature vector was used, comprising statistical moments (mean, standard deviation, skewness, kurtosis), signal range (minimum, maximum, peak-to-peak amplitude), Hjorth parameters (activity, mobility, complexity), spectral moment estimates (spectral centroid and spectral spread), spectral band powers in the theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz) bands, zero-crossing rate, and mean absolute first difference. The inclusion of band-power features follows common EEG feature engineering practice; in particular, beta-band activity has been associated with cognitive control, working memory, and the maintenance of current sensorimotor or cognitive states [31,32]. The RBF kernel was tuned using grid search over $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1\}$ using only the training partition.
2. 1D-CNN: A one-dimensional convolutional neural network with Conv64–Conv128–Conv256 layers followed by global average pooling, trained from scratch.
3. ResNet-1D [7]: Residual blocks adapted for raw one-dimensional time-series classification.
4. InceptionTime [8]: Inception-style modules applied to raw one-dimensional signals.
5. ROCKET [9]: Random convolutional kernels followed by a ridge classifier, using 10,000 kernels as in the original method. ROCKET was applied directly to raw one-dimensional time series without image conversion.
6. RP + InceptionV3 [10]: Recurrence-plot image encoding followed by InceptionV3 fine-tuning.
7. MTF + InceptionV3 [3]: Markov Transition Field encoding using $Q = 4$ quantile bins, followed by InceptionV3 fine-tuning.
8. Spectrogram + InceptionV3 [33]: Short-time Fourier transform magnitude image using a Hann window of 64 points and 75% overlap. Frequency bins were clipped to the interval $[0, \text{Nyquist}/2]$, and the magnitude image was normalised to $[0, 255]$ for CNN input.
9. SAX + BoW + InceptionV3 [13]: Symbolic Aggregate approXimation with Bag-of-Words histogram representation, rasterised into a fixed-size image. Alphabet size $a = 4$ and word length $w = 4$ were used.
10. GASF + InceptionV3 [3]: Gramian Angular Summation Field image encoding followed by InceptionV3 fine-tuning.
11. GADF + InceptionV3 [3]: Gramian Angular Difference Field image encoding followed by InceptionV3 fine-tuning.

12. CGAF + InceptionV3: The proposed Composite Gramian Angular Field encoding followed by the same InceptionV3 fine-tuning protocol used for the other image-based methods.

For PTB-DB, VGG16 [34] results are additionally reported to provide contextual comparison with Elmir et al. [4]. These results should not be interpreted as a strict head-to-head comparison unless the data selection, preprocessing, splitting, and training protocol are identical.

Fairness Controls

All image-based methods used the same InceptionV3 fine-tuning protocol: convolutional layers frozen for 10 epochs, followed by 20 epochs of full fine-tuning using Adam with learning rate $\eta = 0.001$ and batch size 32. Thus, performance differences among image-based methods are primarily attributable to the image encoding rather than to changes in the CNN backbone or training schedule.

4.5. Metrics

The primary evaluation metric was macro-averaged F1 score, because several datasets contained class imbalance. Accuracy, precision, and recall are also reported [35]. For repeated cross-validation experiments, results are reported as mean $\pm$ standard deviation over 30 fold-repetition scores, corresponding to 10 folds repeated three times. For UCR datasets using the standard training–test split, final accuracy is reported on the held-out test partition.

4.6. Statistical Analysis

Pairwise statistical comparisons between CGAF and competing methods were performed using Wilcoxon signed-rank tests [36] over the fold-level scores. The tables report the pairwise Wilcoxon p -values (p_W) obtained for each comparison with CGAF. Where multiple pairwise comparisons were interpreted jointly, Holm–Bonferroni correction [37] was applied at $\alpha = 0.05$ to control the family-wise error rate.

Because repeated cross-validation folds are not fully independent, statistical tests over fold-level scores should be interpreted cautiously. This concern is consistent with earlier analyses of approximate statistical tests for comparing supervised learning algorithms [38]. Following the recommendations of Nadeau and Bengio [39], corrected variance estimates were used where applicable to reduce the optimism associated with overlapping training sets. In addition to p -values, Cohen’s d was reported as an effect-size measure, with practical significance interpreted together with statistical significance.

Friedman and Post Hoc Analysis

A Friedman test was used as a supplementary omnibus analysis across methods. The analysis was treated as descriptive support rather than as the sole basis for inference, because fold-repetition scores are not independent datasets. A standard Nemenyi critical-difference diagram was not used for the domain-specific repeated-CV results, since the Nemenyi test is designed for comparisons across independent datasets and would be misleading if applied to fold-repetition scores as if they were independent datasets. Pairwise comparisons are therefore reported using Wilcoxon signed-rank tests, with Holm–Bonferroni correction used when interpreting families of comparisons.

4.7. Data Leakage Prevention

All model-selection and representation-selection steps were performed within the training data available for each fold. In particular, candidate CGAF weights were selected

using only the training and validation portions of the fold, and the held-out test partition was not used during weight optimisation, hyperparameter selection, CNN fine-tuning decisions, or baseline model selection.

For EEG, subject-level grouping was enforced before cross-validation so that segments from the same participant could not appear in both training and test partitions. For PTB-DB, the outer folds were constructed at the heartbeat-segment level. The absence of patient identifiers in the distributed files prevented patient-level grouping.

For FordA and the UCR datasets, the predefined UCR training–test split was retained, and all optimisation was performed within the training partition. Sample-wise transformations, such as per-sample min–max normalisation and polar embedding, were applied identically to training and test samples and did not use class labels. Learned or selected parameters, including CGAF weights and classifier hyperparameters, were estimated only from training data.

5. Results

5.1. Main Classification Results on Domain-Specific Datasets

This subsection reports the results on the three domain-specific datasets: EEG cognitive engagement, PTB-DB ECG, and FordA automotive fault detection. For EEG, results are reported over subject-level stratified 10-fold cross-validation repeated three times. For PTB-DB, results are reported over heartbeat-segment-level stratified 10-fold cross-validation repeated three times; patient-level grouping could not be performed because patient identifiers were unavailable. For FordA, cross-validation on the training partition was used for model and weight selection, and the final CGAF accuracy is additionally reported on the standard UCR held-out test partition. In the tables, p_W denotes the pairwise Wilcoxon signed-rank test p -value for comparison with CGAF, and d denotes Cohen's effect size. Multiple-comparison interpretation is handled as described in Section 4.6.

5.1.1. EEG Cognitive Engagement

Table 6 reports the EEG pilot-study results. CGAF achieved the highest mean performance among the evaluated methods, with macro-F1 = 0.867 ± 0.027 . Because this dataset contains 500 segments from 100 participants and approximately 50 test samples per fold, the results should be interpreted as pilot evidence rather than as population-level generalisation.

Table 6. Classification performance on the EEG cognitive-engagement dataset (pilot study; $N = 500$, 100 participants). Results are reported as mean $\pm$ standard deviation over 10-fold cross-validation repeated three times. Best values are shown in bold.

Method	Accuracy	Precision	Recall	F1	p_W	d
SVM	0.618 ± 0.040	0.624 ± 0.039	0.615 ± 0.041	0.612 ± 0.041	<0.001	2.1
1D-CNN	0.707 ± 0.047	0.712 ± 0.046	0.703 ± 0.048	0.701 ± 0.048	<0.001	1.6
ResNet-1D	0.738 ± 0.043	0.744 ± 0.042	0.735 ± 0.044	0.733 ± 0.043	<0.001	1.4
InceptionTime	0.753 ± 0.041	0.759 ± 0.040	0.750 ± 0.042	0.748 ± 0.041	<0.001	1.3
ROCKET	0.761 ± 0.038	0.767 ± 0.037	0.757 ± 0.039	0.754 ± 0.038	<0.001	1.2
RP	0.794 ± 0.034	0.799 ± 0.033	0.791 ± 0.035	0.789 ± 0.035	0.002	0.9
MTF	0.811 ± 0.031	0.816 ± 0.030	0.808 ± 0.032	0.805 ± 0.032	0.008	0.8
SAX + BoW	0.571 ± 0.044	0.578 ± 0.043	0.569 ± 0.045	0.565 ± 0.044	<0.001	2.3
Spectrogram	0.682 ± 0.043	0.688 ± 0.042	0.679 ± 0.044	0.675 ± 0.043	<0.001	2.0
GASF	0.817 ± 0.030	0.822 ± 0.029	0.814 ± 0.031	0.812 ± 0.031	0.015	0.7
GADF	0.833 ± 0.028	0.838 ± 0.027	0.829 ± 0.029	0.827 ± 0.029	0.031	0.6
CGAF	0.871 ± 0.026	0.876 ± 0.025	0.868 ± 0.027	0.867 ± 0.027	—	—

5.1.2. PTB-DB Heartbeat Classification

Table 4 reports the heartbeat-segment-level PTB-DB results. CGAF with InceptionV3 achieved the highest mean performance among the evaluated InceptionV3-based image encodings, with macro-F1 = 0.941 ± 0.018 . This estimate does not represent patient-independent performance. Results from Elmir et al. [4] are included only as contextual reference because differences in data selection, preprocessing, splitting, and fine-tuning protocol may affect comparability.

5.1.3. FordA Automotive Fault Detection

Table 7 reports FordA results. Cross-validation results were computed on the UCR training partition and were used for method and weight selection. The final CGAF test accuracy was evaluated once on the standard UCR held-out test partition after the weights had been frozen.

Table 7. Classification performance on FordA. Cross-validation columns report mean $\pm$ standard deviation over 10-fold cross-validation repeated three times on the UCR training partition. The final test accuracy is reported on the standard held-out UCR test partition for CGAF after CV-based weight selection. Best cross-validation values are shown in bold.

Method	Accuracy	Precision	Recall	F1	p_w	d	Test Acc
SVM	0.674 ± 0.039	0.670 ± 0.040	0.671 ± 0.041	0.668 ± 0.040	<0.001	2.4	—
1D-CNN	0.792 ± 0.033	0.788 ± 0.034	0.789 ± 0.034	0.785 ± 0.034	<0.001	1.9	—
ResNet-1D	0.872 ± 0.026	0.869 ± 0.027	0.870 ± 0.027	0.867 ± 0.027	<0.001	1.3	—
InceptionTime	0.900 ± 0.022	0.897 ± 0.023	0.898 ± 0.023	0.895 ± 0.023	0.006	0.8	—
ROCKET	0.903 ± 0.019	0.901 ± 0.020	0.902 ± 0.020	0.899 ± 0.019	0.042	0.6	—
RP	0.786 ± 0.034	0.783 ± 0.035	0.783 ± 0.035	0.780 ± 0.035	<0.001	1.9	—
MTF	0.828 ± 0.028	0.824 ± 0.029	0.825 ± 0.029	0.821 ± 0.029	<0.001	1.6	—
SAX + BoW	0.553 ± 0.042	0.549 ± 0.043	0.551 ± 0.043	0.546 ± 0.043	<0.001	3.1	—
GASF	0.867 ± 0.024	0.864 ± 0.025	0.865 ± 0.025	0.862 ± 0.025	0.028	0.7	—
GADF	0.863 ± 0.026	0.860 ± 0.027	0.861 ± 0.027	0.858 ± 0.027	0.031	0.6	—
Spectrogram	0.791 ± 0.032	0.788 ± 0.033	0.789 ± 0.033	0.785 ± 0.033	<0.001	1.9	—
CGAF	0.912 ± 0.020	0.909 ± 0.021	0.910 ± 0.021	0.907 ± 0.021	—	—	0.912

The FordA result is also included in the UCR benchmark table (Table 8) using the same standard held-out test partition. Thus, the FordA domain-specific result and the FordA UCR entry are two presentations of the same final held-out test evaluation, not independent test-set evaluations.

Table 8. Test accuracy on 20 selected UCR datasets. CGAF, GASF, and GADF use the same InceptionV3 backbone and fine-tuning protocol and therefore form the primary GAF-based image-encoding comparison. ROCKET is included as a strong non-image time-series reference baseline. Bold values indicate the highest accuracy in each row. The symbol [†] marks datasets where CGAF is also the best-performing method when the ROCKET reference baseline is included. FordA* is also reported in the domain-specific setup using the same standard held-out test partition.

Dataset	GASF	GADF	CGAF	ROCKET
ECG5000 [†]	0.932	0.935	0.943	0.937
ElectricDevices	0.651	0.648	0.661	0.749
FaceAll [†]	0.792	0.795	0.821	0.808
GunPoint [†]	0.940	0.943	0.960	0.957
ItalyPowerDemand [†]	0.961	0.959	0.971	0.969
MoteStrain [†]	0.872	0.876	0.889	0.881
NonInvFatalECG_Thorax1 [†]	0.878	0.881	0.904	0.891
OSULeaf [†]	0.762	0.765	0.788	0.776
SonyAIBORobotSurface1 [†]	0.883	0.879	0.901	0.897
SwedishLeaf [†]	0.918	0.921	0.939	0.932
TwoPatterns [†]	0.963	0.961	0.977	0.974
UWaveGestureLibraryAll	0.841	0.843	0.857	0.887
Wafer [†]	0.989	0.988	0.993	0.991
FordA*	0.863	0.858	0.912	0.919
Earthquakes	0.701	0.698	0.714	0.762
FreezerRegularTrain [†]	0.971	0.969	0.982	0.978
HandOutlines	0.872	0.869	0.885	0.907
MiddlePhalanxOutlineAgeGroup	0.612	0.615	0.627	0.651
SmoothSubspace [†]	0.953	0.950	0.967	0.961
Yoga [†]	0.831	0.833	0.851	0.847
CGAF better than GASF and GADF	20/20 datasets			
CGAF best among GAF-based image encodings	20/20 datasets			
CGAF best including ROCKET reference baseline	14/20 datasets			
ROCKET higher than CGAF	6/20 datasets: ElectricDevices, UWaveGestureLibraryAll, FordA, Earthquakes, HandOutlines, MiddlePhalanxOutlineAgeGroup			

5.2. UCR Benchmark Results

Table 8 reports test accuracy of CGAF, GASF, GADF, and ROCKET on the 20 selected UCR datasets. The primary methodological comparison is between CGAF, GASF, and GADF, because these methods belong to the same GAF-based image-encoding family and use the same InceptionV3 fine-tuning protocol. ROCKET is included as a strong non-image time-series reference baseline, but it should not be interpreted as a direct methodological analogue of CGAF. The resulting comparison is not intended as a comprehensive state-of-the-art leaderboard, and no conclusion is drawn about the rank of CGAF relative to transformer-based or foundation-model classifiers that were not evaluated.

For each UCR dataset, CGAF weight optimisation was performed using only the training partition. The final accuracy was then evaluated once on the standard UCR held-out test partition. Therefore, no per-dataset statistical test is reported for the UCR results; the win counts are descriptive summaries over the selected benchmark subset.

FordA appears both in the domain-specific results and in the UCR benchmark table. These entries refer to the same standard UCR held-out test partition and the same final CGAF test accuracy. FordA is counted only once in all UCR win-rate summaries.

Across the selected UCR subset, CGAF outperformed both individual GAF components on all 20 datasets. This is the primary comparison for assessing the proposed representation, because GASF, GADF, and CGAF share the same GAF-based image-encoding

framework and CNN backbone. When ROCKET was included as a strong non-image reference baseline, CGAF achieved the highest accuracy on 14 of 20 datasets, while ROCKET achieved the highest accuracy on six datasets: ElectricDevices, UWaveGestureLibraryAll, FordA, Earthquakes, HandOutlines, and MiddlePhalanxOutlineAgeGroup. These results suggest that CGAF improves the GAF-based image pipeline, while also showing that specialised non-image time-series classifiers may remain preferable for some datasets.

5.3. Weighting Strategy Comparison

Table 9 compares the proposed weight-selection procedure with simpler weighting alternatives. The proposed magnitude-ratio initialisation followed by grid search and Bayesian refinement achieved the highest mean F1 on both EEG and PTB-DB. These results support the use of data-driven weight selection, while the comparison should be interpreted as an ablation of the weighting strategy rather than as an exhaustive optimisation study.

Table 9. F1 score for different weighting strategies using InceptionV3. Results are reported as mean $\pm$ standard deviation over the corresponding cross-validation runs. The final column reports the difference in EEG F1 relative to the proposed full optimisation procedure.

Strategy	EEG F1	PTB-DB F1	Δ_{EEG}
Fixed ($w_s = w_d = 0.5$)	0.791 $\pm$ 0.034	0.897 $\pm$ 0.031	−0.075
Entropy-based	0.841 $\pm$ 0.031	0.922 $\pm$ 0.022	−0.025
Variance-ratio	0.844 $\pm$ 0.029	0.919 $\pm$ 0.024	−0.022
Magnitude-ratio + Bayesian	0.867 $\pm$ 0.027	0.941 $\pm$ 0.018	—

5.4. Ablation: Sensitivity to Fusion Parameter λ

To separate the effect of the relative GASF–GADF balance from the effect of unconstrained weight magnitude, we evaluated a one-parameter constrained version of CGAF:

$$G_{\lambda}(i, j) = (1 - \lambda) \cos(\phi_i + \phi_j) + \lambda \sin(\phi_i - \phi_j), \quad \lambda \in [0, 1]. \quad (15)$$

This corresponds to $w_s = 1 - \lambda$ and $w_d = \lambda$, with $w_s + w_d = 1$. The constrained formulation is therefore a subset of the full CGAF search space used in the main experiments.

The sensitivity of classification performance to the constrained fusion parameter λ is reported in Table 10.

Table 10. Sensitivity of CGAF performance to the constrained fusion parameter λ . The constrained formulation uses $w_s = 1 - \lambda$ and $w_d = \lambda$, with $w_s + w_d = 1$. The final column reports the full unconstrained CGAF result. Values are reported as mean and standard deviation over the corresponding fold-repetition scores. Best mean values for each dataset are shown in bold.

Dataset/Statistic	Constrained CGAF: λ											Full CGAF
	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.66	0.71	0.80	1.00	Unconstr.
EEG mean	0.812	0.831	0.844	0.855	0.860	0.851	0.857	0.862	0.867	0.851	0.827	0.867
EEG SD	0.030	0.029	0.028	0.028	0.028	0.029	0.028	0.027	0.027	0.028	0.030	0.027
PTB-DB mean	0.881	0.897	0.911	0.921	0.930	0.928	0.937	0.941	0.938	0.922	0.894	0.941
PTB-DB SD	0.031	0.030	0.029	0.027	0.025	0.025	0.023	0.022	0.023	0.025	0.030	0.018

For EEG, the best constrained value occurs near $\lambda = 0.71$, matching the mean F1 of the full CGAF result. For PTB-DB, the best constrained mean occurs near $\lambda = 0.66$, also matching the full CGAF mean F1, while the unconstrained version has a smaller standard deviation in the reported fold-repetition scores. This pattern is consistent with a possible

stability benefit of the unconstrained search in these experiments, but it should not be interpreted as a theoretical guarantee of superiority over the constrained formulation.

5.5. Friedman Omnibus Analysis

Table 11 reports Friedman test statistics for ranking the evaluated methods on the domain-specific datasets. The test is included as a supplementary omnibus analysis. Because the scores come from repeated cross-validation folds rather than from independent datasets, the results should be interpreted cautiously and not used as the sole basis for statistical inference. Pairwise comparisons with CGAF are reported in the main result tables using Wilcoxon signed-rank tests, with multiple-comparison interpretation handled as described in Section 4.6.

Table 11. Supplementary Friedman omnibus analysis across fold-level scores on the domain-specific datasets. The analysis is descriptive because repeated cross-validation folds are not independent datasets.

Dataset	Friedman χ^2	<i>p</i> -Value	CGAF Rank
EEG (pilot study)	87.4	<0.001	1 (best)
PTB-DB	91.2	<0.001	1 (best)
Ford A	76.8	<0.001	1 (best)

A Nemenyi critical-difference diagram was not computed for these domain-specific repeated-CV results. The standard Nemenyi post hoc test is designed for comparisons across independent datasets; applying it to fold-repetition scores as if they were independent datasets would overstate the strength of the evidence.

5.6. AlexNet Cross-Architecture Validation

AlexNet validation was performed on the EEG and PTB-DB datasets to provide a preliminary check of whether the observed CGAF advantage over GASF and GADF was specific to the InceptionV3 backbone. FordA and UCR AlexNet results were not included because of computational constraints; full cross-architecture validation across all evaluated datasets is left for future work.

The AlexNet cross-architecture results are reported in Table 12.

Table 12. AlexNet validation on EEG and PTB-DB. Results are reported as mean $\pm$ standard deviation over the corresponding cross-validation runs. The same fold indices and weight-optimisation protocol were used as in the InceptionV3 experiments.

Method	EEG Acc	EEG F1	PTB-DB Acc	PTB-DB F1
GASF	0.768 $\pm$ 0.038	0.764 $\pm$ 0.039	0.869 $\pm$ 0.032	0.865 $\pm$ 0.032
GADF	0.783 $\pm$ 0.036	0.779 $\pm$ 0.037	0.881 $\pm$ 0.030	0.877 $\pm$ 0.030
CGAF	0.825 $\pm$ 0.031	0.820 $\pm$ 0.032	0.923 $\pm$ 0.025	0.919 $\pm$ 0.025

The AlexNet results are consistent with the InceptionV3 experiments in that CGAF outperforms the individual GASF and GADF encodings on the two evaluated datasets. However, because only two datasets and one additional backbone are included, this should be interpreted as preliminary cross-architecture support rather than as comprehensive architecture-independent validation.

5.7. Weight Stability Across Folds

Table 2 reports the optimised CGAF weights and clipping fractions for the three domain-specific datasets. The fold-level standard deviations of the selected weights are

small in these experiments, indicating that the optimisation procedure selected similar weight regions across cross-validation folds. This supports, but does not prove, the stability of the weight-selection procedure.

Across the three domain-specific datasets, the optimised difference-component weight w_d was larger than the summation-component weight w_s . This pattern is reported descriptively and should not be interpreted as a universal property of CGAF. It may reflect dataset-specific angular structure, but establishing such an interpretation would require additional controlled experiments.

The clipped-pixel fraction was below 0.5% for all three datasets, indicating that the fixed $[-2, 2]$ clipping boundary affected only a small fraction of pixels for the optimised weights used here. For future datasets or manually selected weights with larger $w_s + w_d$, the clipping range should be reported explicitly and, if necessary, adjusted to the corresponding raw CGAF range.

5.8. Ablation Study

Table 13 reports an ablation study on the EEG pilot dataset. The aim is to separate the contribution of the individual GAF components, the weight-optimisation procedure, and the selected weight range. The results should be interpreted as an ablation of the proposed pipeline on one representative dataset rather than as a universal statement about all possible time-series domains.

Table 13. Ablation study on the EEG pilot dataset using InceptionV3. Results are reported as mean $\pm$ standard deviation F1 over 10-fold cross-validation repeated three times.

Variant	EEG F1	Δ vs. CGAF
CGAF (full; magnitude-ratio init. + grid + Bayesian)	0.867 $\pm$ 0.027	—
<i>Trigonometric component ablation</i>		
Cosine only ($w_d = 0$; equivalent to GASF)	0.812 $\pm$ 0.030	−0.054
Sine only ($w_s = 0$; equivalent to GADF)	0.827 $\pm$ 0.029	−0.039
Equal weights ($w_s = w_d = 0.5$)	0.791 $\pm$ 0.034	−0.075
<i>Weight optimisation strategy ablation</i>		
Magnitude-ratio init. only (no grid, no Bayesian)	0.836 $\pm$ 0.030	−0.030
Coarse grid only (no Bayesian refinement)	0.858 $\pm$ 0.028	−0.008
Bayesian only (random start, no grid init.)	0.849 $\pm$ 0.031	−0.017
Grid + Bayesian (full, as proposed)	0.867 $\pm$ 0.027	—
<i>Weight range ablation</i>		
$w_s, w_d \in [0, 1]$	0.847 $\pm$ 0.029	−0.019
$w_s, w_d \in [0, 2]$ (proposed)	0.866 $\pm$ 0.027	—
$w_s, w_d \in [0, 5]$	0.831 $\pm$ 0.033	−0.035

The component ablation indicates that both individual GAF components perform below the full CGAF representation on the EEG pilot dataset. The equal-weight configuration also performs below the optimised CGAF variant, suggesting that the relative scaling of the summation and difference components is important for this dataset. However, this result should not be interpreted as evidence that equal weighting is generally harmful; it only shows that equal weighting was suboptimal under the present EEG protocol.

The optimisation-strategy ablation shows that coarse grid search already captures most of the observed gain, while Bayesian refinement provides a small additional improvement in mean F1. The weight-range ablation suggests that the proposed range $[0, 2] \times [0, 2]$ was adequate for the EEG experiment: limiting the range to $[0, 1]$ reduced performance, whereas expanding it to $[0, 5]$ did not improve results and increased variability.

Combining the optimisation-strategy ablation with the measured runtime provides an explicit efficiency–accuracy trade-off. Magnitude-ratio initialisation without CNN-based weight search achieved EEG macro-F1 = 0.836. Coarse-grid optimisation increased macro-F1 to 0.858 at a measured cost of approximately 37 GPU-minutes per fold. Adding Bayesian refinement increased the total weight-selection time to approximately 45 GPU-minutes and improved macro-F1 to 0.867. Therefore, Bayesian refinement required approximately eight additional GPU-minutes, corresponding to a 22% increase relative to the grid-only runtime, for an absolute macro-F1 improvement of 0.009. Grid-only optimisation may consequently provide a reasonable lower-cost configuration when computational resources are constrained, whereas the complete grid-plus-Bayesian procedure is appropriate when maximising validation performance is prioritised. This trade-off was measured on the EEG pilot dataset and should not be assumed to generalise unchanged to other datasets or CNN backbones.

5.9. Computational Complexity Analysis

The computational cost of CGAF has two distinct components: the cost of encoding a time series after the weights have been selected, and the one-time training cost of selecting the weights. These two costs should be interpreted separately. At inference time, CGAF produces a single image and therefore has the same output memory footprint as a single GAF encoding. During training, weight optimisation requires additional proxy CNN evaluations.

The encoding and weight-optimisation costs are summarised in Table 14.

At inference time, CGAF has a runtime similar to computing GASF and GADF separately because both trigonometric components must be evaluated. Its storage advantage concerns the resulting representation: CGAF retains one composite image rather than two separate GAF images. The stored output is approximately 0.08 MB for CGAF and approximately 0.16 MB when GASF and GADF images are retained separately. The current implementation constructs both component matrices as temporary intermediate arrays; consequently, these output-storage values should not be interpreted as measured peak process RAM.

Table 14. Computational complexity and encoding runtime per sample ($m = 100$, Intel i7-9700K, CPU only, single-threaded, batch size 1). “Separate” denotes computing and storing GASF and GADF as two matrices. Runtime is reported as mean $\pm$ standard deviation over 1000 repeated measurements. Stored-output size denotes the memory required for the resulting encoded image or images and excludes temporary intermediate arrays. Weight optimisation is a training-time cost and is not part of inference.

Method/Step	Time Complexity	Space	Stored Output (MB)	Runtime
GASF only	$\mathcal{O}(m^2)$	$\mathcal{O}(m^2)$	~ 0.08	3.1 ± 0.2 ms
GADF only	$\mathcal{O}(m^2)$	$\mathcal{O}(m^2)$	~ 0.08	3.1 ± 0.2 ms
GASF + GADF (separate)	$\mathcal{O}(2m^2)$	$\mathcal{O}(2m^2)$	~ 0.16	6.2 ± 0.3 ms
CGAF encoding	$\mathcal{O}(m^2)$	$\mathcal{O}(m^2)$	~ 0.08	5.9 ± 0.3 ms
<i>Weight optimisation cost; training only, not inference</i>				
Grid search (7×7 , per fold, GPU)	$\mathcal{O}(49 \cdot N_{tr} m^2)$	$\mathcal{O}(m^2)$	—	~ 37 min/fold
Bayesian refinement (20 evaluations, per fold, GPU)	—	—	—	~ 8 min/fold
Total weight optimisation	—	—	—	~ 45 min/fold

The reported weight-optimisation cost is a training-time cost. It is incurred once per fold or train–validation split and is not part of inference-time classification. In the present implementation, the full weight search required approximately 45 GPU-minutes per fold on an NVIDIA RTX 2080 Ti. This additional cost should be considered when applying

CGAF to new datasets, especially when large numbers of folds, datasets, or backbone architectures are evaluated.

The measurements in Table 15 quantify only runtime scaling with m . Classification accuracy was not measured systematically for $m \in \{50, 150, 200\}$; therefore, the runtime results should not be interpreted as an accuracy-sensitivity analysis.

Table 15. CGAF encoding runtime as a function of the PAA target length m (CPU only, single-threaded, mean $\pm$ standard deviation over 1000 repeated measurements). The observed growth is consistent with the expected $\mathcal{O}(m^2)$ scaling.

m	50	100	150	200	Measured ratio	Theoretical m^2 ratio
Runtime (ms)	1.6 ± 0.1	5.9 ± 0.3	13.1 ± 0.5	23.4 ± 0.8	1:3.69:8.19:14.63	1:4:9:16

Overall, CGAF should be described as memory-efficient at inference relative to storing separate GASF and GADF images but not as computationally free. The encoding cost remains quadratic in the target length m , and the weight-optimisation stage adds a non-negligible training-time overhead.

6. Discussion

6.1. Why Did CGAF Improve Performance in the Evaluated Settings?

CGAF exposes two structurally different angular relationships to the same convolutional filters. GASF contributes symmetric summation-based patterns, including diagonal self-relations and morphology-related pairwise structure, whereas GADF contributes anti-symmetric difference-based patterns that preserve the ordering and direction of temporal changes. Their weighted integration can therefore be useful when class discrimination depends jointly on signal morphology and directional dynamics, and neither component alone provides a sufficient image representation.

This interpretation is consistent with the EEG ablation in Table 13: cosine-only, sine-only, and equal-weight variants all performed below the fully optimised CGAF representation. The result suggests that both component inclusion and their relative scaling contributed under the evaluated protocol. Across the selected UCR datasets, CGAF also outperformed the individual GASF and GADF encodings under the same CNN backbone, supporting the usefulness of joint representation within the tested subset.

CGAF should not be expected to provide the same benefit for every signal. Its advantage may be smaller when one component is redundant or noise-dominated, when extensive padding or compression obscures the original temporal structure, or when discriminative long-range patterns are represented more effectively by a specialised one-dimensional classifier. The cases in which ROCKET outperformed CGAF are consistent with this possibility, although the present experiments do not establish a definitive mapping between dataset characteristics and the preferred classifier.

6.2. UCR Benchmark Interpretation

The primary UCR comparison is within the GAF-based image-encoding family because CGAF, GASF, and GADF use the same InceptionV3 backbone and differ principally in their image representation. CGAF outperformed both individual components on all 20 selected datasets, supporting the empirical value of their weighted integration within this subset. ROCKET was included only as a strong one-dimensional reference baseline and achieved the highest accuracy on six datasets. These results indicate that CGAF improves the evaluated GAF-image pipeline but does not replace specialised one-dimensional classi-

fiers. A full UCR/UEA evaluation and dataset-level meta-analysis would be required to determine systematically which signal properties favour each approach.

6.3. Interpretation of Learned Weights

The selected weights should be interpreted as dataset-dependent representation-selection parameters rather than as direct physical or physiological indicators. A larger w_d assigns greater emphasis to directional angular differences, whereas a larger w_s emphasises summation-based symmetric structure. Weights outside the unit-sum constraint also permit amplification of one or both components. The observed values indicate which mixture improved validation performance under each experimental protocol, but they do not establish a universal relationship between particular signal domains and optimal CGAF weights. CGAF is therefore an optimised representation family rather than a fixed parameter-free encoding.

6.4. Explainability and Component Attribution

Future work should investigate the classification decisions of CGAF-based CNNs using explainable-AI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM). Such analysis could identify spatial regions of the CGAF image that contribute most strongly to individual classification decisions. Because each CGAF pixel is an algebraic combination of weighted GASF and GADF values, a Grad-CAM map alone cannot determine which component produced a salient region. Component attribution should therefore combine Grad-CAM with analysis of the corresponding regions in $w_s G_{\text{GASF}}$ and $w_d G_{\text{GADF}}$, together with controlled component ablation or occlusion. This would help determine whether particular classification decisions are driven primarily by summation-based angular relationships, difference-based directional relationships, or their interaction.

6.5. Limitations

Several limitations should be considered when interpreting the results.

1. **Benchmark and comparative scope.** The 20 evaluated UCR datasets constitute a selected, heterogeneous subset rather than a formally stratified sample of the complete UCR/UEA archive. The primary controlled comparison is between CGAF, GASF, and GADF under the same CNN backbone. ROCKET is included as a strong one-dimensional reference and outperformed CGAF on six datasets. PatchTST, TimesNet, MOMENT, and other transformer-based or foundation-model classifiers were not evaluated. Consequently, the results should not be interpreted as a comprehensive state-of-the-art ranking.
2. **EEG pilot-study constraints.** The EEG dataset contains 500 segments from 100 primary-school pupils and was recorded using a consumer-grade single-channel device. Although subject-level cross-validation was used, the cohort and the number of test samples per fold remain limited. Labels were based on behavioural observation and task performance rather than direct neurophysiological ground truth, and the implemented filtering and amplitude gating provide pragmatic rather than clinical-grade artefact rejection. The EEG findings should therefore be interpreted as pilot evidence only.
3. **PTB-DB evaluation granularity.** The preprocessed PTB-DB files do not contain patient identifiers. Repeated cross-validation was consequently performed at the heartbeat-segment level, and heartbeat samples originating from the same patient may have appeared in different folds. The reported PTB-DB performance does not establish patient-independent or clinical generalisation.

4. Dependence of repeated cross-validation scores. Fold-repetition scores are not fully independent because their training partitions overlap. Standard deviations, statistical tests, and effect-size estimates based on these scores should therefore be interpreted cautiously and not as substitutes for external validation.
5. PAA target length and image-size sensitivity. Classification-accuracy sensitivity to the PAA target length m was not systematically evaluated. The fixed value $m = 100$ was applied uniformly across the compared image encodings, but it is not necessarily optimal for every dataset or encoding. Runtime scaling for $m \in \{50, 100, 150, 200\}$ is reported in Table 15, but it does not replace an accuracy-sensitivity analysis. For short datasets such as SmoothSubspace ($L = 15$), padding also creates a large homogeneous image region. Future work should compare fixed, adaptive, and dataset-specific target lengths.
6. Backbone dependence and optimisation cost. Most image-based experiments use InceptionV3. The additional AlexNet evaluation provides only preliminary cross-architecture evidence. Moreover, the complete grid-plus-Bayesian weight-selection procedure requires approximately 45 GPU-minutes per fold. Grid-only optimisation provides a lower-cost alternative, but broader evaluation across efficient CNN backbones and lightweight weight-selection procedures remains necessary.
7. Representation scope. CGAF is an optimised representation family rather than a fixed parameter-free encoding, whereas GASF and GADF are evaluated as fixed individual representations. CGAF also uses only a first-order linear fusion and is defined here for univariate time series. Non-linear fusion, multichannel formulations, and multivariate extensions may provide additional benefits but require separate methodological and empirical evaluation.

7. Conclusions

This paper introduced CGAF, a Composite Gramian Angular Field representation for time-series classification. CGAF combines the Gramian Angular Summation Field (GASF) and Gramian Angular Difference Field (GADF) components into a single weighted image before CNN training. The weights were selected using coarse grid search followed by Gaussian-process Bayesian refinement. All weight selection was restricted to training-only inner validation partitions, and the selected weights were frozen before held-out test evaluation.

The structural analysis shows that the summation and difference GAF components are algebraically non-redundant trigonometric functions. This observation motivates their combination but does not prove that CGAF contains more discriminative information or will improve classification. The principal evidence for the method is therefore empirical and applies only to the evaluated datasets, architectures, and experimental protocols.

Across the three domain-specific datasets, CGAF improved over the individual GASF and GADF encodings in the reported experiments. It achieved macro-F1 = 0.867 on the EEG cognitive-engagement pilot study, heartbeat-segment-level macro-F1 = 0.941 on PTB-DB, and test accuracy = 91.2% on FordA. The PTB-DB result did not establish patient-level generalisation because patient identifiers were unavailable. On the selected subset of 20 UCR datasets, CGAF outperformed both GASF and GADF on all 20 datasets under the same InceptionV3 image-classification pipeline. When ROCKETS was included as a strong non-image reference baseline, CGAF achieved the highest accuracy on 14 of 20 datasets, while ROCKETS achieved the highest accuracy on the remaining six.

The results support CGAF as a useful extension of the GAF image-encoding family. Its main advantage is that it integrates summation-based and difference-based angular information into a single image without adding gradient-trained parameters or modifying

the CNN backbone. At inference time, CGAF has the same output memory footprint as a single GAF image, although it still requires evaluation of both trigonometric components and therefore should not be described as computationally free.

The EEG results should be interpreted as pilot evidence only. Broader validation requires larger EEG cohorts, additional recording devices, external validation, and comparison across additional CNN and non-CNN backbones. Future work should also evaluate CGAF on the full UCR/UEA archive, test modern architectures such as EfficientNet and transformer-based time-series models, extend the method to multivariate signals, and investigate lighter-weight strategies for transferring or approximating the CGAF weight-selection procedure.

Author Contributions: Conceptualisation, P.B. and S.M.; methodology, P.B. and S.M.; software, P.B.; validation, P.B. and S.M.; formal analysis, P.B.; investigation, P.B. and S.M.; data curation, S.M.; writing—original draft preparation, P.B.; writing—review and editing, S.M. and A.G.; supervision, S.M. and A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The EEG data used in this study were collected during school-based science popularisation activities involving primary-school pupils. The activities used the NeuroSky MindWave device, a consumer-grade brain–computer interface device intended for educational, gaming, and demonstration purposes, and not a medical diagnostic device. Participation in the activities was organised according to the usual school practice for extracurricular and science-popularisation events. Written parental consent for children’s participation was obtained before the activities, and parents were informed about the purpose and nature of the activity. During the activity, children received feedback related to task-related attention and engagement indicators while solving tasks. The present study used anonymised records derived from those previously conducted educational activities as secondary data. No medical diagnosis, clinical intervention, or health-related decision-making was performed. The authors did not process directly identifiable personal data for the present analysis. Because the data were anonymised and used retrospectively for methodological research on time-series classification, no additional individual consent was sought for the present analysis. All procedures were conducted in accordance with applicable data-protection principles and with respect for the privacy and welfare of the participating children.

Informed Consent Statement: Written parental consent was obtained before the school-based science-popularisation activities in which the EEG recordings were collected. The present study used anonymised secondary data derived from those activities. No directly identifiable personal data were used in the analysis reported in this paper.

Data Availability Statement: The preprocessed PTB-DB heartbeat files used in this study are publicly available through the Kaggle ECG Heartbeat Categorization Dataset (<https://www.kaggle.com/datasets/shayanfazeli/heartbeat>; accessed on 30 June 2026) [29]. They were derived from the PTB Diagnostic ECG Database [28]. The FordA dataset and the UCR benchmark datasets used in this study are publicly available from the UCR Time Series Classification Archive [30]. The EEG dataset used in this study is not publicly available because it was derived from school-based activities involving minors, and public release could create privacy and data-protection risks even after anonymisation. Aggregated results, preprocessing details, fold-level protocols, and implementation information are available from the corresponding author upon reasonable request, subject to applicable ethical, institutional, and data-protection constraints.

Acknowledgments: The authors acknowledge the institutional context provided by the project “Artificial Intelligence for Future Technical Education and Industrial Competitiveness (AITECH), IP-UNIST-48”. The project is funded by the European Union–NextGenerationEU through the National Recovery and Resilience Plan 2021–2026. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. During the

preparation of this manuscript, the authors used ChatGPT (GPT-5.5 Thinking, OpenAI) for language editing, improvement of scientific wording, and assistance with LaTeX/MDPI formatting. The authors reviewed and edited the generated output and take full responsibility for the content of this publication.

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

Appendix A. Preprocessing Pipeline

This appendix summarises the preprocessing steps used before GAF or CGAF image construction. Throughout all experiments, min–max normalisation (Equations (1) and (2)) was applied to each original-length time series before PAA resampling to the target length $m = 100$. The same preprocessing procedure was applied consistently to training, validation, and test samples within each dataset. Parameters that required selection or fitting, such as CGAF weights or classifier hyperparameters, were selected using training data only.

Appendix A.1. EEG Preprocessing

The EEG data were recorded using the NeuroSky MindWave device during school-based science-popularisation activities. Because this is a consumer-grade single-channel device rather than a clinical EEG system, the preprocessing pipeline should be interpreted as pragmatic artefact reduction for methodological time-series classification, not as clinical-grade EEG processing.

The input data consisted of 3 s single-channel EEG segments recorded at the FP1 position with a sampling frequency of 512 Hz. The implemented preprocessing comprised the following steps:

1. Band-pass filtering from 0.5 to 50 Hz using a zero-phase fifth-order Butterworth filter.
2. Segment-level amplitude gating. A segment was rejected in its entirety when the maximum absolute amplitude of the filtered signal exceeded 100 μV . This criterion was intended to reject large-amplitude transient contamination, including prominent frontal blink artefacts, although lower-amplitude ocular contamination may remain.
3. Per-sample min–max normalisation, mapping to $[-1, 1]$, followed by PAA resampling to $m = 100$ before CGAF encoding.

Class labels were assigned using behavioural observation and task-performance information independently of the EEG preprocessing and CGAF features.

Appendix A.2. PTB-DB Heartbeat Data

The PTB heartbeat samples were obtained from the preprocessed ECG Heartbeat Categorization Dataset distributed through Kaggle. The authors did not process the original multi-lead PhysioNet recordings and did not perform lead selection, QRS detection, band-pass filtering, notch filtering, or heartbeat segmentation. The supplied fixed-length heartbeat samples were used as distributed [29] and were loaded from `ptbdb_abnormal.csv` and `ptbdb_normal.csv`, min–max normalised per sample, mapped to $[-1, 1]$, and resampled to $m = 100$ using PAA before CGAF encoding.

Appendix A.3. FordA and UCR Preprocessing

For FordA and the UCR datasets, no additional domain-specific filtering was applied. Each time series was min–max normalised to $[0, 1]$, mapped to $[-1, 1]$, and resampled to $m = 100$ using PAA before GAF or CGAF encoding. The standard UCR training–test split was retained where available, and all weight-selection steps were performed using only the training partition.

Appendix B. Reproducibility Details

Appendix B.1. Software Environment

- Python 3.9.13.
- TensorFlow 2.11.0 (GPU-enabled).
- scikit-learn 1.2.0/scipy 1.10.0/numpy 1.23.5.
- scikit-optimize 0.9.0 (Bayesian optimisation).
- CUDA 11.8/cuDNN 8.6.

Appendix B.2. Hardware

- CPU: Intel Core i7-9700K; GPU: NVIDIA RTX 2080 Ti (11 GB).
- RAM: 32 GB DDR4-3200; OS: Ubuntu 22.04 LTS.

Appendix B.3. Hyperparameters

- Batch size: 32; Adam $\eta = 0.001$ ($\beta_1 = 0.9$, $\beta_2 = 0.999$).
- Epochs: 10 frozen-backbone epochs followed by 20 full fine-tuning epochs.
- Learning-rate scheduling: ReduceLROnPlateau.
- Grid search proxy: five epochs per candidate; final training uses the full 30-epoch protocol.
- Image resolution: 100×100 pixels, replicated to three channels for CNN input.
- CGAF weight grid: $\{0.1, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0\}^2$ ($7 \times 7 = 49$ candidates).
- Bayesian refinement: 20 evaluations in a local neighbourhood of ± 0.15 .
- Random seed: 42; repeat seeds: $\{42, 43, 44\}$.
- EEG evaluation: subject-level stratified 10-fold cross-validation repeated three times; weight selection used a fixed subject-level inner training–validation split within each outer training fold.
- PTB-DB evaluation: heartbeat-segment-level stratified 10-fold cross-validation repeated three times; weight selection used a fixed stratified 80/20 training–validation split within each outer training fold.
- UCR weight optimisation: performed independently for each dataset using only the corresponding training partition.

References

1. Craik, A.; He, Y.; Contreras-Vidal, J.L. Deep Learning for EEG Classification Tasks: A Review. *J. Neural Eng.* **2019**, *16*, 031001. <https://doi.org/10.1088/1741-2552/ab0ab5>.
2. Lotte, F.; Bougrain, L.; Cichocki, A.; Clerc, M.; Congedo, M.; Rakotomamonjy, A.; Yger, F. A Review of EEG-Based BCI Classification Algorithms: 10-year Update. *J. Neural Eng.* **2018**, *15*, 031005. <https://doi.org/10.1088/1741-2552/aab2f2>.
3. Wang, Z.; Oates, T. Imaging Time-Series to Improve Classification and Imputation. *arXiv* **2015**, <https://doi.org/10.48550/arXiv.1506.00327>.
4. Elmir, Y.; Himeur, Y.; Amira, A. ECG Classification Using Deep CNN and Gramian Angular Field. In Proceedings of the 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService), Athens, Greece, 17–20 July 2023; pp. 137–141. <https://doi.org/10.1109/BigDataService58306.2023.00026>.
5. Hatami, N.; Gavet, Y.; Debayle, J. Classification of Time-Series Images Using Deep CNNs. In *Tenth International Conference on Machine Vision (ICMV 2017)*; SPIE: Bellingham, WA, USA, 2018; *arXiv* **2015**, <https://doi.org/10.1117/12.2309486>.
6. Zhao, Q.; Zhao, D.; Yin, W. EEG Emotion Recognition Based on GADF and AMB-CNN Model. *Int. J. Numer. Model. Electron. Netw. Devices Fields* **2024**, *37*, e70000. <https://doi.org/10.1002/jnm.70000>.
7. Wang, Z.; Yan, W.; Oates, T. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 1578–1585. <https://doi.org/10.1109/IJCNN.2017.7966039>.
8. Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. InceptionTime: Finding AlexNet for Time Series Classification. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962. <https://doi.org/10.1007/s10618-020-00710-y>.

9. Dempster, A.; Petitjean, F.; Webb, G.I. ROCKET: Exceptionally Fast and Accurate Time Series Classification Using Random Convolutional Kernels. *Data Min. Knowl. Discov.* **2020**, *34*, 1454–1495. <https://doi.org/10.1007/s10618-020-00701-z>.
10. Eckmann, J.-P.; Kamphorst, S.O.; Ruelle, D. Recurrence Plots of Dynamical Systems. *Europhys. Lett.* **1987**, *4*, 973–977. <https://doi.org/10.1209/0295-5075/4/9/004>.
11. Garcia-Ceja, E.; Uddin, M.Z.; Torresen, J. Classification of Recurrence Plots Distance Matrices with CNN'. *Procedia Comput. Sci.* **2018**, *130*, 157–163. <https://doi.org/10.1016/j.procs.2018.04.025>.
12. Mallat, S.G. A Theory for Multiresolution Signal Decomposition. *IEEE Trans. PAMI* **1989**, *11*, 674–693. <https://doi.org/10.1109/34.192463>.
13. Lin, J.; Keogh, E.; Wei, L.; Lonardi, S. Experiencing SAX: A Novel Symbolic Representation of Time Series. *Data Min. Knowl. Discov.* **2007**, *15*, 107–144. <https://doi.org/10.1007/s10618-007-0064-z>.
14. Barra, S.; Carta, S.M.; Corrigan, A.; Podda, A.S.; Recupero, D.R. Deep Learning and Time Series-to-Image Encoding for Financial Forecasting. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 683–692. <https://doi.org/10.1109/JAS.2020.1003132>.
15. Yoon, G.-W.; Joo, S. Enhanced electrocardiogram classification using Gramian Angular Field transformation with multi-lead analysis and segmentation techniques. *MethodsX* **2025**, *14*, 103297. <https://doi.org/10.1016/j.mex.2025.103297>.
16. Vargas, V.M.; Ayllón-Gavilán, R.; Durán-Rosal, A.M.; Gutiérrez, P.A.; Hervás-Martínez, C.; Guijo-Rubio, D. Gramian Angular and Markov Transition Fields for Ordinal Classification. In *Advances in Computational Intelligence; IWANN 2023*; Springer: Cham, Switzerland, 2023; Volume 14135. https://doi.org/10.1007/978-3-031-43078-7_41.
17. Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C.K.; Li, X. Self-Supervised Contrastive Representation Learning for Semi-Supervised Time-Series Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11170–11185. <https://doi.org/10.1109/TPAMI.2023.3308189>.
18. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *arXiv* **2022**. <https://doi.org/10.48550/arXiv.2211.14730>.
19. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis'. *arXiv* **2022**. <https://doi.org/10.48550/arXiv.2210.02186>.
20. Goswami, M.; Szafer, K.; Choudhry, A.; Cai, Y.; Li, S.; Dubrawski, A. MOMENT: A Family of Open Time-series Foundation Models'. *arXiv* **2024**. <https://doi.org/10.48550/arXiv.2402.03885>.
21. Feofanov, V.; Wen, S.; Alonso, M.; Ilbert, R.; Guo, H.; Tiomoko, M.; Pan, L.; Zhang, J.; Redko, I. Mantis: Lightweight Calibrated Foundation Model for User-Friendly Time Series Classification. *arXiv* **2025**, arXiv:2502.15637. <https://doi.org/10.48550/arXiv.2502.15637>.
22. Zhang, H.; Liu, Y.; Qiu, Y.; Liu, H.; Pei, Z.; Wang, J.; Long, M. TimesBERT: A BERT-Style Foundation Model for Time Series Understanding. *arXiv* **2025**, arXiv:2502.21245. <https://doi.org/10.48550/arXiv.2502.21245>.
23. Yeh, C.-C.M.; Saini, U.S.; Wang, J.; Dai, X.; Fan, X.; Sun, J.; Fan, Y.; Zheng, Y. TiCT: A Synthetically Pre-Trained Foundation Model for Time Series Classification. *arXiv* **2025**, arXiv:2511.19694. <https://doi.org/10.48550/arXiv.2511.19694>.
24. Liu, Z.; Wang, Y.; Li, B.; Zheng, J.; Eldele, E.; Wu, M.; Ma, Q. A Unified Shape-Aware Foundation Model for Time Series Classification. *arXiv* **2026**, arXiv:2601.06429. <https://doi.org/10.48550/arXiv.2601.06429>.
25. Xu, C.; Tao, D.; Xu, C. A Survey on Multi-view Learning. *arXiv* **2013**, arXiv:1304.5634. <https://doi.org/10.48550/arXiv.1304.5634>.
26. Keogh, E.; Chakrabarti, K.; Pazzani, M.; Mehrotra, S. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowl. Inf. Syst.* **2001**, *3*, 263–286. <https://doi.org/10.1007/PL00011669>.
27. Snoek, J.; Larochelle, H.; Adams, R. Practical Bayesian Optimization of Machine Learning Algorithms. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS) 25*, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 2951–2959. <https://doi.org/10.48550/arXiv.1206.2944>.
28. Bousseljot, R.; Kreiseler, D.; Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomed. Tech.* **1995**, *40*, S317–S318. <https://doi.org/10.1515/bmte.1995.40.s1.317>.
29. Kachuee, M.; Fazeli, S.; Sarrafzadeh, M. ECG Heartbeat Classification: A Deep Transferable Representation. In *Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI)*, New York, NY, USA, 4–7 June 2018; pp. 443–444. <https://doi.org/10.1109/ICHI.2018.00092>.
30. Bagnall, A.; Lines, J.; Bostrom, A.; Large, J.; Keogh, E. The Great Time Series Classification Bake Off. *Data Min. Knowl. Discov.* **2017**, *31*, 606–660. <https://doi.org/10.1007/s10618-016-0483-9>.
31. Engel, A.K.; Fries, P. Beta-Band Oscillations—Signalling the Status Quo?. *Curr. Opin. Neurobiol.* **2010**, *20*, 156–165. <https://doi.org/10.1016/j.conb.2010.02.015>.
32. Schmidt, R.; Ruiz, M.H.; Kilavik, B.E.; Lundqvist, M.; Starr, P.A.; Aron, A.R. Beta Oscillations in Working Memory, Executive Control, and Sensorimotor Function. *J. Neurosci.* **2019**, *39*, 8231–8238. <https://doi.org/10.1523/JNEUROSCI.1163-19.2019>.
33. Allen, J.B. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1977**, *25*, 235–238. <https://doi.org/10.1109/TASSP.1977.1162950>.

34. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**. <https://doi.org/10.48550/arXiv.1409.1556>.
35. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>.
36. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. <https://doi.org/10.2307/3001968>.
37. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70. **1979** <https://www.jstor.org/stable/4615733>.
38. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. <https://doi.org/10.1162/089976698300017197>.
39. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Mach. Learn.* **2003**, *52*, 239–281. <https://doi.org/10.1023/A:1024068626366>.

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