Enhancing Time Series Classification with Diversity-Driven Neural Network Ensembles

Javidan Abdullayev
Université de Haute Alsace
IRIMAS
Mulhouse, France
javidan.abdullayev@uha.fr

Maxime Devanne
Université de Haute Alsace
IRIMAS
Mulhouse, France
maxime.devanne@uha.fr

Cyril Meyer
Université de Haute Alsace
IRIMAS
Mulhouse, France
cyril.meyer@uha.fr

Ali Ismail-Fawaz
Université de Haute Alsace
IRIMAS
Mulhouse, France
ali-el-hadi.ismail-fawaz@uha.fr

Jonathan Weber
Université de Haute Alsace
IRIMAS
Mulhouse, France
jonathan.weber@uha.fr

Germain Forestier
Université de Haute Alsace
IRIMAS
Mulhouse, France
Monash University, DSAI
Melbourne, Australia
germain.forestier@uha.fr

***Abstract*—Ensemble methods have played a crucial role in achieving state-of-the-art (SOTA) performance across various machine learning tasks by leveraging the diversity of features learned by individual models. In Time Series Classification (TSC), ensembles have proven highly effective whether based on neural networks (NNs) or traditional methods like HIVE-COTE. However most existing NN-based ensemble methods for TSC train multiple models with identical architectures and configurations. These ensembles aggregate predictions without explicitly promoting diversity which often leads to redundant feature representations and limits the benefits of ensembling. In this work, we introduce a diversity-driven ensemble learning framework that explicitly encourages feature diversity among neural network ensemble members. Our approach employs a decorrelated learning strategy using a feature orthogonality loss applied directly to the learned feature representations. This ensures that each model in the ensemble captures complementary rather than redundant information. We evaluate our framework on 128 datasets from the UCR archive and show that it achieves SOTA performance with fewer models. This makes our method both efficient and scalable compared to conventional NN-based ensemble approaches.**

***Index Terms*—Time Series Classification, Deep Learning, Ensemble Learning, Decorrelated Learning**

I. Introduction

Time Series Classification (TSC) is a fundamental problem in machine learning with applications across various domains including healthcare [1], human activity recognition [2], social security [3], remote sensing [4], etc.. Increasing availability of large-scale time series datasets, such as the UCR archive [5], has led to the development of more advanced classification methods. Recent progress in deep learning has significantly improved TSC performance by leveraging convolutional neural networks (CNNs) which can automatically extract meaningful temporal features [6].

Currently state-of-the-art (SOTA) deep learning models for TSC including InceptionTime [7], H-InceptionTime [8] and LITETime [9] achieve high accuracy using ensemble learning. The *Time* suffix in their names signifies that these models are ensembles of five identical neural networks, each trained independently. These methods combine predictions from their individual members to improve classification performance. However they do not explicitly enforce feature diversity within the ensemble just relying instead on random initialization to introduce variation. This often leads to feature redundancy and limits the potential gains from ensembling.

In ensemble learning, diversity among individual models is key for improving generalization [10]. Traditional approaches such as HIVE-COTE [11] achieve diversity by ensembling heterogeneous models with different architectures. However, in this work, we take a different approach by focusing on homogeneous neural network ensembles. Our goal is to promote diversity among models with the same architecture by encouraging them to learn distinct feature representations. This is an important but often overlooked aspect of deep ensembles where models trained independently often converge to similar solutions.

To address this issue we propose a decorrelated learning framework that explicitly promotes feature diversity within ensembles by penalizing redundant representations. Inspired by knowledge distillation [12], we introduce a feature orthogonality loss that forces ensemble models to learn complementary rather than overlapping features. This loss minimizes the cosine similarity between feature vectors produced by different ensemble members. As a result, each model captures unique aspects of the input, reducing redundancy across the ensemble. This enhances generalization without increasing computational complexity or modifying the base model architecture.

The overall idea is exemplified using the BirdChicken dataset from the UCR archive, as shown in Figure 1. On the left, two base LITE models trained separately with different initializations produce highly similar feature maps. Conversely, in our proposal, a second decorrelated LITE model is guided to learn diverse features compared to the first base model trained previously. Figure 1 clearly illustrates on the right side that the decorrelated model learned different features. This differentiation enables the decorrelated ensemble to achieve 100% test accuracy which exceeds the 90% accuracy achieved by the base ensemble. Furthermore, it also exceeds the best performance of SOTA models such as InceptionTime [7] which achieves a maximum test accuracy of 95% using an ensemble of five models. It suggests that learning diverse features in an ensemble of deep models can result in better generalization and classification performance.

We assessed our approach on 128 datasets from the UCR archive, demonstrating its effectiveness across diverse real-world time series data. As a result our approach achieves performance comparable to LITETime [9] but with fewer models which offers a more efficient solution.

Our main contributions in this work are:

- We propose a novel diversity-driven ensemble framework that explicitly promotes feature diversity and improves deep ensemble effectiveness for TSC.
- We conduct a comprehensive empricial validation across 128 UCR datasets, showing that our method achieves SOTA-level performance with fewer models and improves efficiency.
- We provide quantitative and qualitative diversity analysis to demonstrate increased feature diversity using Fréchet Inception Distance (FID) scores and t-SNE visualizations of learned convolutional filters.

The rest of the paper is organized as follows: Section II provides background information and discusses related work on deep learning-based TSC and ensemble methods. Section III describes our proposed decorrelated learning framework. Section IV presents the experimental setup, dataset details and evaluation metrics, followed by a detailed analysis of results. Finally Section V concludes the paper and discusses potential future directions.

II. Background and Related Work

In this work we focus on TSC task where recent studies demonstrated that deep neural networks (DNNs) leveraging 1D temporal convolutions achieve impressive performance [13]. Many traditional TSC algorithms especially those not based on deep learning rely heavily on feature engineering as part of the classification task. While this approach has shown strong performance in certain cases [6] it often requires domain expertise and poses challenges in scalability and automation. On the other hand deep learning models integrate feature extraction and classification into a single pipeline and optimize both jointly. This capability allows them to scale efficiently to large datasets and take advantage of hardware acceleration such as GPUs.

SOTA deep learning models for TSC often employ ensemble learning where multiple models of the same architecture are trained independently with different initializations and their predictions are combined [9]. While ensembling has been shown to improve accuracy [10], this approach does not explicitly enforce feature diversity among individual models. As a result, ensembles may suffer from redundant feature representations which limits their overall effectiveness.

In this work, we address this issue by focusing on increasing diversity among ensemble members. Specifically, we introduce a decorrelated feature loss that actively encourages diverse feature learning across models during training. Our approach optimizes ensemble diversity by ensuring that individual models learn complementary instead of redundant representations which leads to improved generalization and classification performance.

A. Definitions

A univariate time series is defined as an ordered set of numerical values that represents evolution of a specific quantity over time. A time series dataset is denoted as $\mathcal{D} = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^{N}$ where N represents the number of samples, $\mathbf{X}_i$ is an individual time series and $\mathbf{Y}_i$ is its corresponding label vector. The label vector follows a one-hot encoding scheme where $\mathbf{Y}_i \in \mathbb{R}^C$ represents a class label c from a set of C predefined categories.

Time series classification (TSC) task involves assigning a class label to a given sample based on its temporal characteristics. The objective is to train a model that can effectively identify patterns, trends and dependencies within time series data. Formally, the task consists of learning a mapping function $f : \mathbf{X} \rightarrow \mathbf{Y}$ that accurately classifies each input time series $\mathbf{X}_i$ into one of the predefined categories in $\mathbf{Y}_i$.

B. Deep Learning for Time Series Classification

Time series analysis has been a fundamental area of research for many years with various machine learning techniques applied to tasks like TSC. Early approaches often relied on similarity-base methods such as Dynamic Time Warping (DTW) [14] between time series or classification models such as Random Forests [15] and Support Vector Machines (SVMs) [16]. A key limitation of these traditional methods is that feature extraction is typically treated as a separate process from classification which can lead to information loss and increase the complexity of the overall pipeline.

In recent years deep learning has gained significant popularity in the TSC due to its ability to learn complex patterns from raw time series data without requiring manual feature engineering. One of the first deep learning models applied to TSC was the Multi-Layer Perceptron (MLP) [17] but its fully connected nature made it inefficient in capturing temproal dependencies. A major step forward came with the introduction of 1D CNNs which proved highly effective for extracting meaningful features from time series data. The Fully Convolutional Network (FCN) [18] was one of the first deep learning models to achieve strong results in TSC. FCN consists of three convolutional blocks, each containing a convolutional

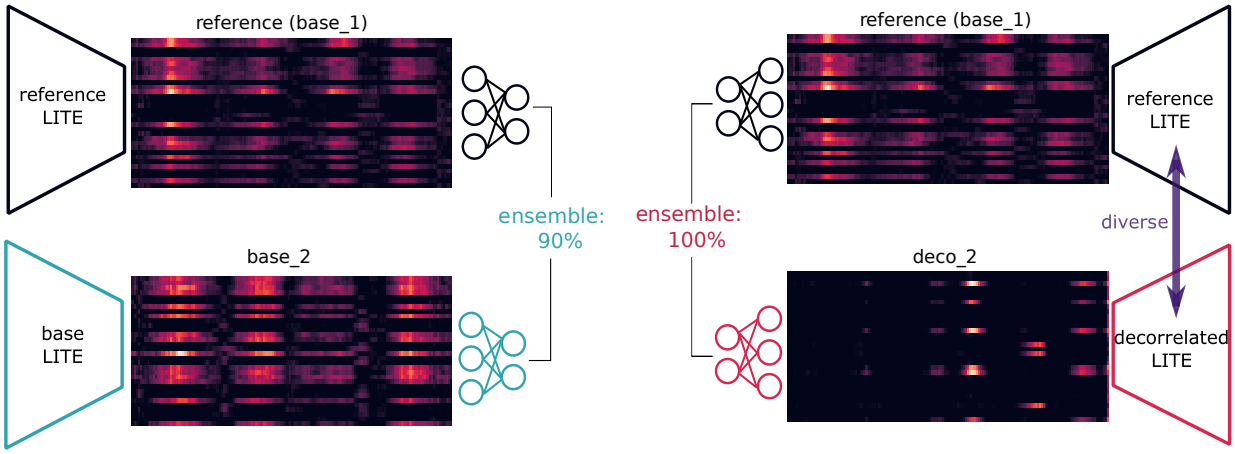

Fig. 1. Comparison of ensemble model performances and feature maps on the BirdChicken dataset, from standard and decorrelated training.

layer, batch normalization and activation functions. Unlike standard CNN architectures, FCN does not use pooling layers which allows it to preserve the original time series length and retain temporal relationships more effectively. Following FCN, researchers introduced ResNet for TSC [18] which applies nine convolutional layers but also incorporates residual connections. These connections improve gradient flow, mitigate information loss and make training deeper models more stable. In 2020, Inception-based model were introduced for TSC in the form of InceptionTime [7], drawing inspiration from Google's Inception v4 [19]. InceptionTime consists of six Inception modules where each module applies convolutions of different kernel sizes to capture patterns at varying temporal resolutions. Building on InceptionTime, an improved model called Hybrid Inception (H-Inception) [8] was developed, incorporating custom convolutional filters in the initial layers to enhance feature extraction and classification performance.

Recently, LITE (Light Inception with Boosting Techniques) was proposed as a more efficient alternative to Inception-based model [9]. LITE significantly reduces parameter count to just 2.34% of that of InceptionTime while maintaining competitive classification performance. The model consists of three convolutional layers, combining custom, multiplexed, dilated and depthwise separable convolutions to reduce computational cost while preserving predictive accuracy

To achieve SOTA performance, ensemble versions of these models, InceptionTime, H-InceptionTime and LITETime, are widely used in TSC. Each model is trained five times under the same setup but with different initializations and their predictions are combined to produce a final ensemble output. While ensembling improves classification accuracy these approaches do not explicitly enforce diversity among the individual classifiers. Instead they rely solely on random initialization to introduce variation. However, this does not guarantee diverse feature learning and often leads to redundant representations.

C. Ensemble Learning

Ensemble learning is a powerful machine learning technique that improves performance by combining predictions from multiple individual models. It leverages their collective strengths to mitigate individual weaknesses and reduce generalization errors [10]. Common ensemble strategies include bagging where models are trained on different subsets of the training data [20], boosting which iteratively focuses on difficult-to-classify samples [21] and stacking where predictions of base models are combined through a meta-learner [22].

In deep learning, ensemble approaches have been highly successful in improving generalization and reducing overfitting, particularly in complex tasks such as image recognition and time series classification. The work [17] demonstrated the effectiveness of deep neural network ensembles for time series classification by highlighting the importance of combining diverse predictions for improved accuracy. The recent model CoCaLite [23] which utilizes ensemble strategies to balance computational efficiency and predictive accuracy has demonstrated SOTA performance. It is important to note that most existing works overlook critical aspects of feature diversity within ensembles. Various techniques such as decorrelated learning and feature orthogonality have been introduced to explicitly promote diversity during training [24]. These methods are particularly relevant for deep neural networks where similar architectures and training setups can lead to redundant feature extraction and limits ensemble benefits. For instance, the kernels tailored for time series similarity have been employed with ensemble models to enhance temporal pattern recognition [25].

The key advantage of ensembling lies in its ability to exploit the diversity of features learned by each individual model. Because of the stochastic nature of deep learning, each model may learn slightly different features which collectively enhance overall performance of the ensemble. In convolutional neural networks this diversity is largely attributed to the variation in features learned by each model. Our experiments

have shown that when the models in an ensemble learn very similar features, the performance gains are minimal. This observation is really intuitive and underscores the importance of feature diversity in the success of ensemble learning.

D. Decorrelated Learning

To the best of our knowledge, most SOTA deep learning-based TSC models rely on training multiple instances of the same model and ensembling their outputs [7]–[9]. Base models within these ensembles are trained independently without mechanisms for sharing information or coordinating their learning. This independent training process often leads the model to converge similar, nearby local minima which results in reduced diversity. Both theoretical and experimental studies suggest that generalization ability of an ensemble can be greatly enhanced if base models are negatively correlated [26].

In contrast, decorrelated learning explicitly introduces shared training mechanisms that encourage diversity among ensemble members. Instead of training models independently, decorrelated learning ensures that training process of each model is influenced by others in the ensemble. By explicitly minimizing correlations between filters, features or predictions decorrelated learning ensures that each model is guided to learn distinct patterns. This approach helps the features produced by different models to complement one another rather than overlap. This shared training paradigm is particularly valuable in ensemble settings where diversity among individual models plays a curcial role in improving overall performance [27].

Techniques such as Orthogonality Loss have been developed to enforce feature decorrelation by minimizing cosine similarity between filters [28]. These methods are particularly relevant for tasks where redundant feature extraction limits performance, as is often observed in deep neural networks trained on datasets with identical architectures. Most existing approaches in CNN-based models apply orthogonality loss to convolutional filters to achieve diversity in the feature space [29].

In this work, we demonstrate that for TSC tasks applying orthogonality loss to convolutional filters does not lead to meaningful diversity in features. To address this, we propose applying orthogonality loss directly to the feature outputs. This approach ensures a higher level of feature diversity which is critical for improving ensemble performance. Additionally, decorrelation has been shown to be effective in unsupervised representation learning, helping to mitigate representational collapse and improve downstream task performance [24]. These findings reinforce the value of decorrelated learning in addressing redundancy and enhancing the discriminative power of deep learning models.

Our method promotes diversity explicitly at the level of convolutional features, a strategy we have found effective for ensemble learning in TSC tasks. We utilize a diversity-driven auxiliary loss to improve diversity among ensemble models during training. The core idea of our decorrelated learning approach is to train ensemble models sequentially, ensuring that each model learns features distinct from those learned by previously trained models. In this paper, we focus on exploring the impact of decorrelated learning in convolution-based TSC models to deepen our understanding of its effects.

III. Proposed Method

In this section we introduce our diversity-driven ensemble learning framework for TSC. We first describe the transition from filter-level orthogonality to feature-level orthogonality, highlighting its advantages in promoting meaningful diversity among ensemble members. Next we provide a detailed description of our proposed framework, outlining the sequential training process and incorporation of feature orthogonality constraints. Finally we present the mathematical formulation of the feature orthogonality loss which serves as the core of our method.

A. Rethinking Diversity: From Filter Orthogonality to Feature Orthogonality

The primary objective of this work is to enhance diversity among ensemble models while maintaining the performance of individual models. One possible approach to promote feature diversity is to use techniques such as Deep Negative Correlation Classification (DNCC) [30] where diversity is encouraged during training by explicitly forcing one model to focus on samples that are misclassified or have low confidence in another model. However, in our experiments with LITE model [9] we observed that it achieves 100% training accuracy on majority of datasets from UCR archive This indicates that the model effectively fits training data and leaves minimal room for misclassified or low-confidence samples to enhance diversity through misclassification-based methods. Given this limitation we explored alternative methods for promoting diversity, focusing on feature and filter-level diversity. From the literature we identified several works that employ filter orthogonality loss to promote diversity among ensemble models, particularly in image classification tasks [28], [29], [31]. The central idea behind these methods is to enforce orthogonality between convolutional filters, encouraging them to learn distinct representations. From our experiments we realized that filter orthogonality often shifts discriminative patterns rather than creating diverse representations. While filters become orthogonal, feature maps remain similar, limiting the benefits of decorrelation. Thus filter-level orthogonality alone does not guarantee meaningful diversity in ensembles of time series classifiers.

To address this challenge, we propose a feature diversity loss function and a diversity-driven optimization strategy aimed at encouraging each model in the ensemble to extract distinct features from the input data. Unlike existing approaches that target filter diversity our method focuses on promoting diversity directly in the feature space itself. Specifically, we introduce an orthogonality loss function that operates on the feature outputs. The goal is to explicitly enforce orthogonality between the feature representations produced by different models in the ensemble. This loss function encourages each model to learn

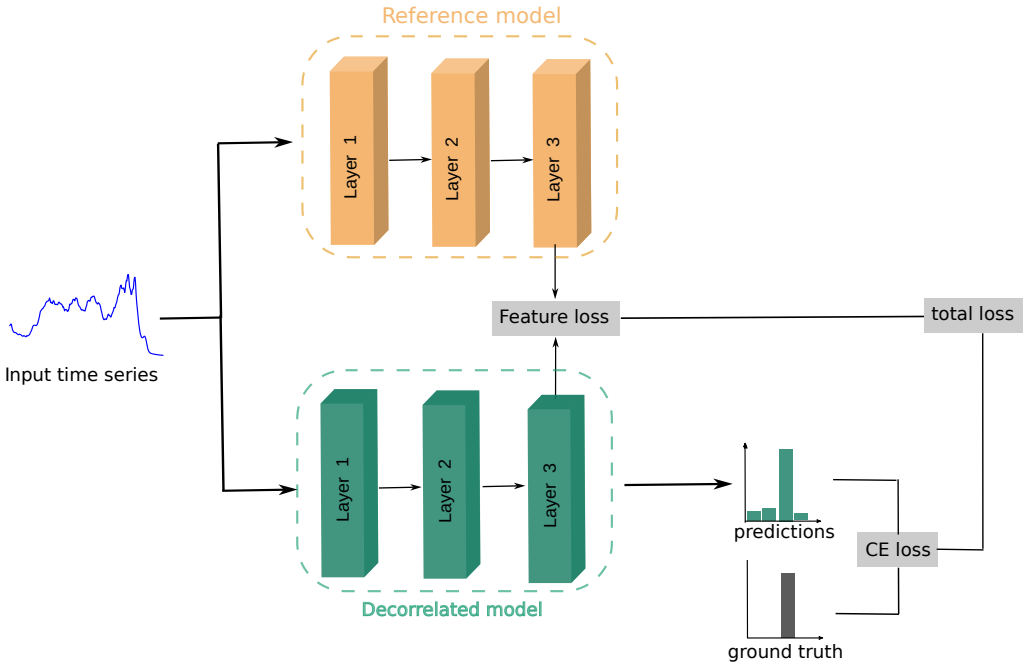

Fig. 2. Proposed decorrelated learning framework for a 2-model ensemble where the decorrelated model is trained with feature orthogonality loss to enhance diversity.

orthogonal features thereby reducing redundancy in the feature space. By encouraging orthogonal feature representations we aim to promote diversity among individual models within the ensemble, ultimately improving the generalization and overall performance. This approach ensures that each model contributes unique information to the ensemble, leading to a more robust and generalizable classification system.

B. Framework Overview

In this work we introduce a sequential training framework designed to enhance diversity in neural network ensembles for TSC. Our approach explicitly encourages each model in the ensemble to learn distinct feature representations thereby improving generalization. Unlike conventional ensemble methods where models are trained independently our framework introduces decorrelated models which are explicitly guided to learn features that are orthogonal to those of previously trained models. The training follows a sequential process where each base model in the ensemble (n) is trained to minimize feature redundancy with respect to all earlier models $(< n)$.

To achieve this goal we define an orthogonality function that measures the degree of overlap between the feature representations of the current model and those of previously trained models. This function penalizes feature similarity, ensuring that the newly trained model captures complementary information. The resulting feature orthogonality loss is combined with cross-entropy loss to form the training loss. By optimizing both losses jointly, our framework promotes feature diversity while maintaining high classification accuracy.

Figure 2 illustrates the overall architecture of our proposed method, highlighting how cross-entropy (CE) loss and feature orthogonality (FO) loss are integrated into the training process. CE loss ensures that each model learns to accurately map time series inputs to class labels. Meanwhile FO loss is applied directly to the feature outputs of the final convolutional layer to encourage diversity among ensemble members.

A key consideration in DNNs is that lower layers typically learn generic features while upper layers capture task-specific representations [32]. Applying orthogonality constraints to early layers can disrupt essential feature learning, negatively impacting performance. To mitigate this issue we apply FO loss only to the final layer of the LITE model, ensuring that only high-level features are decorrelated thereby maintaining both diversity and classification performance.

The total loss function balances CE loss and FO loss, ensuring that both classification accuracy and feature diversity are optimized during training. In ensemble configurations, this process is iteratively applied to all base models. For each new model FO loss is computed relative to all previously trained models and the results are averaged to ensure consistency. During training, feature outputs from the base models are extracted and each decorrelated model is explicitly optimized to produce feature representations that are distinct from its predecessors. This approach enhances ensemble diversity, leading to a more robust and generalizable classification system while maintaining performance.

C. Diversity Loss

To enhance diversity among ensemble models, we employ a feature orthogonality loss based on cosine similarity, applied directly to the feature outputs. While traditional methods focus on filter-level orthogonality, we found that encouraging orthog-

onality in the feature space leads to more meaningful diversity in the context of time series data. This ensures that each model in the ensemble contributes unique, complementary information.

Let $\mathbf{F}_{\text{deco}} \in \mathbb{R}^{B\times C\times T}$ and $\mathbf{F}_{\text{base}} \in \mathbb{R}^{B\times C\times T}$ represent the feature outputs of the decorrelated and base models where B, C and T denote batch size, number of channels, and time series length, respectively. The diversity loss is defined as in the following equation 1:

$$L_{orth} = \sum_{i\neq j} \left| \frac{\mathbf{F}_{deco,i} \cdot \mathbf{F}_{base,j}^{\top}}{|\mathbf{F}_{deco,i}||\mathbf{F}_{base,j}|} \right| \tag{1}$$

This formulation penalizes overlap between features, encouraging decorrelated model to learn distinct representations relative to the base models. By penalizing the magnitude of off-diagonal elements in the similarity matrix, we encourage feature independence while maintaining computational efficiency. Notably, the use of cosine similarity is not computationally expensive as matrix operations can be efficiently parallelized. The diversity loss is then integrated into the total loss function as shown in the equation 2:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{CE}} + (1-\alpha)\cdot\mathcal{L}_{\text{orth}} \tag{2}$$

$\mathcal{L}_{\text{CE}}$ is the cross-entropy loss which ensures accurate classification and α is a weight parameter that balances cross-entropy and feature diversity. In our experiments, we set $\alpha = 0.5$, giving equal importance to both losses. By minimizing this total loss, the decorrelated model learns features that are both task-relevant and orthogonal to those of previously trained models, enhancing diversity across the ensemble.

IV. Experimental Evaluation

A. Experimental Setup

1) Data: To validate effectiveness of our proposed approach we evaluated it on UCR Archive [5], the largest publicly available repository for time series classification. The archive includes 128 univariate time series datasets from various domains, such as healthcare, motion tracking and sensor data, with diverse characteristics in terms of sequence length, sample size, and class distribution. The number of classes ranges from 2 to 60, providing a broad evaluation spectrum. For a fair comparison with state-of-the-art methods we used the original train/test splits provided by the UCR Archive. All time series were z-normalized to ensure zero mean and unit variance, reducing the influence of scale differences and emphasizing intrinsic temporal patterns.

2) Experimental protocol: For our experiments, we followed the exact same setup as described in the LITE paper [9]. We trained five standard LITE classifiers and also four decorrelted LITE classifiers which we refer to as *base* and *decorrelated* models respectively. The *base* models trained using only CE loss while the decorrelated models incorporated an additional feature diversity loss. To ensure a robust comparison and confirm that performance changes are not due to random initialization, each base model was initializaed with a unique seed. For consistency, decorrelated models were initlized with the same seeds as their corresponding base models. This guarantees that any observed performance improvements are due to the feature orthogonality (FO) loss rather than randomness in initialization.

The training process for decorrelated ensembles can be summarized as follows:

1) *Base Model Training*: The first model in each ensemble is trained using only CE loss.
2) *Decorrelated Model Training*: Subsequent models in the ensemble are trained as decorrelated models.
3) *Sequential Training*: For each additional model, FO loss is computed on all previously trained models from the ensemble. Specifically, the FO loss formula for (n+1)-th model is given by equation 3.

$$L_{\text{orth}}^{n+1} = \frac{1}{n}\sum_{i=1}^{n} L_{\text{orth}}(F_{n+1}, F_i) \tag{3}$$

where F_{n+1} and F_i represent the feature outputs of the $(n+1)$-th and i-th models, respectively, L_{orth} denotes the orthogonality loss and n is the number of previously trained models from the ensemble.

We evaluate four ensemble cofigurations with sizes varying from two to five models and systematically comparing decorrelated ensembles against base enesembles. Each ensemble consists of a reference model combined with either additional independently trained base models or decorrelated models trained sequentially. For base ensembles, all models are trained independently without any form of coordinated learning and their predictions are aggregated. In contrast decorrelated ensembles are constructed by progressively replacing base models with their corresponding decorrelated versions which are trained sequentially. In this process, each decorrelated model is guided to learn feature representations that are explicitly diverse from those of previously trained models in the ensemble. In all cases, the first base model serves as a fixed reference to ensure consistency across configurations. Each ensemble configuration is trained five times independently and the final results are obtained by averaging performance across these five runs. To denote different ensemble types we use LiteTime-N to represent an ensemble of N independently trained base models while Deco-LiteTime-N refers to an ensemble where all but the reference model are decorrelated.

3) Comparison Protocol: We evaluate classification performance using accuracy across the 128 datasets in the UCR archive. To compare our decorrelated framework with standard ensembles, we use the *Multi-Comparison Matrix (MCM)* [33], which ranks classifiers by *Mean Accuracy*—the average accuracy across all datasets. This provides a more interpretable comparison than traditional average-rank methods [34]. MCM also reports the *Mean Difference*, which measures the average accuracy gap between pairs of classifiers, and the *Win/Tie/Loss* counts across datasets. Statistical significance is assessed using the *Wilcoxon signed-rank test* [35] with $p < 0.05$. Significant results are shown in bold.

4) Implementation details: The LITE model was used as the base architecture, following the original configuration described in prior work [9]. The models were trained using the Adam optimizer with an initial learning rate of 0.001, a reducing factor of 0.5 and a patience of 50. Each model was trained for 1500 epochs with a batch size of 64. All experiments were conducted on a system equipped with an NVIDIA RTX 4090 GPU with 24GB of memory, running Ubuntu 22. The models were implemented using PyTorch 2.5.1 and Python 3.12. The source code is publicly available https://github.com/MSD-IRIMAS/decorrelated-learning .

B. Overall Performance on UCR Archive

In this section we present a comparative analysis of our proposed framework against the SOTA LITETime [9]. Our objective is to evaluate effectiveness of our diversity-driven approach in enhancing generalization and classification accuracy by analyzing various ensemble configurations.

The comparative results are presented in the form of an MCM matrix as illustrated in Figure 4. Among all configurations, the 4-model decorrelated ensemble (Deco-LITETime-4) achieves the highest mean accuracy even surpassing the state-of-the-art LITETime-5 which is an ensemble of five LITE classifiers. This performance clearly demonstrate efficiency of decorrelated learning in extracting complementary features which enables superior classification accuracy with fewer models. The p-value between the Deco-LITETime-4 and LITETime-5 indicates that these two classifiers are not statistically different. In contrast the p-value between LITETime-4 and LITETime-5 is lower than 5% which highlights a statistically significant difference. From the figure we can also notice that Deco-LITETime-4 also outperform its couterpart, LITETime-4 with a statistically significant lower p-value. Furthermore, the results from the Figure 4 also indicate that the Deco-LITETime-4 performs almost on par with LITETime-5 despite using one fewer model. Additionally, a one-vs-one performance comparison between Deco-LITETime-4 and LITETime-5 is shown in the Figure 3.

Similarly, from the Figure 4, we can also observe that the 3-model decorrelated ensemble (Deco-LITETime-3) achieves comparable performance to the more parameter-intensive LITETime-5. The difference between Deco-LITETime-3 and LITETime-5 is not statistically significant, as indicated by the high p-value. The performance of Deco-LITETime-3 establishes it as a viable alternative to the SOTA LITETime-5 while offering comparable results with two fewer models.

From the same figure, it seems that advantages of decorrelated learning are less evident in smaller ensembles. The 2-model decorrelated ensemble (Deco-LITETime-2) shows only a marginal improvement over its base counterpart (LITETime-2) and does not fuly bridge performance gap with LITETime-5. These results shows that benefits of decorrelated learning become more apparent as ensemle size increases. It can be explained based on the fact that, in smaller ensembles, features from 2-base models already exhibit some inherent diversity which diminishes immediate impact of decorrelation. The feature overlap among base models increases as the ensemble size grows in which decorrelated learning plays a crucial role in maximizing feature diversity and enahncing generalization and overall performance.

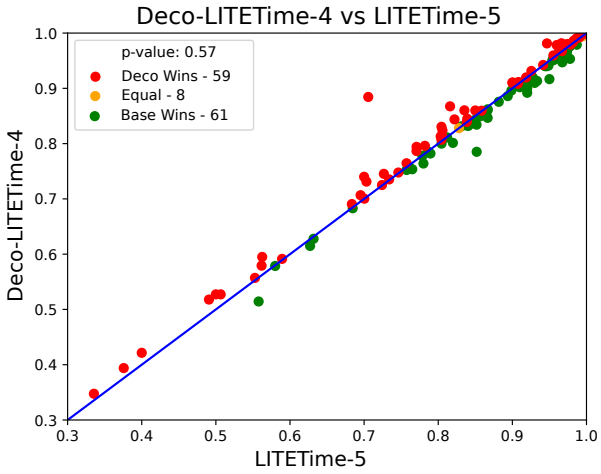

Fig. 3. Performance comparison between Deco-LITETime-4 and LITETime-5.

C. Comparison with State-of-the-Art Methods

As shown in Figure 5, our proposed method achieves an average accuracy of 0.8496, ranking third among the evaluated models. It performs competitively against other SOTA deep learning approaches and surpasses several existing methods. The only approach that significantly outperforms it with statistical significance is MultiROCKET [36] which remains the highest-ranked model. This highlights the effectiveness of our diversity-driven ensemble strategy in improving classification performance while maintaining efficiency. Notably, the Deco-LITETime-4 contains less than 10% of the parameters of a single Inception-based classifier, demonstrating that our method achieves strong performance with significantly reduced computational cost. It is important to note that while the results for our baseline were obtained through our own experiments, the results for the other methods, including MultiROCKET, were sourced from the official repository [36]. These findings further demonstrate that explicitly promoting feature diversity within ensembles enhances generalization and provides a strong alternative to conventional ensemble learning techniques.

D. Quantitative Diversity Analysis

To quantitatively assess the impact of our decorrelated learning framework on feature diversity, we compare two different ensemble configurations using the Fréchet Inception Distance (FID) [37], [38]. FID measures the similarity between two distributions where each distribution represents the statistical properties of features extracted by an individual model. By computing FID scores within two distinct 2-model ensembles, we can directly compare the feature diversity of base models

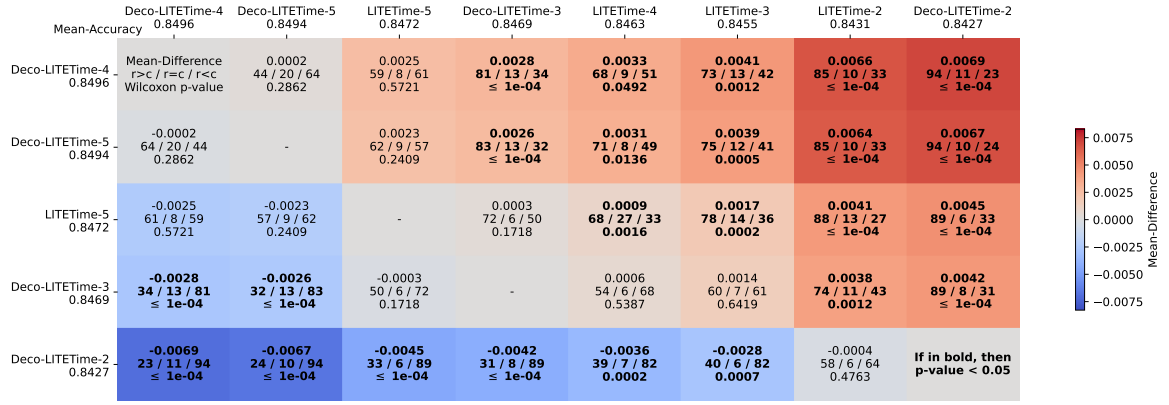

Fig. 4. The Multi-Comparison Matrix illustrates performance of each decorrelated and base ensemble variants in one-vs-one comparisons.

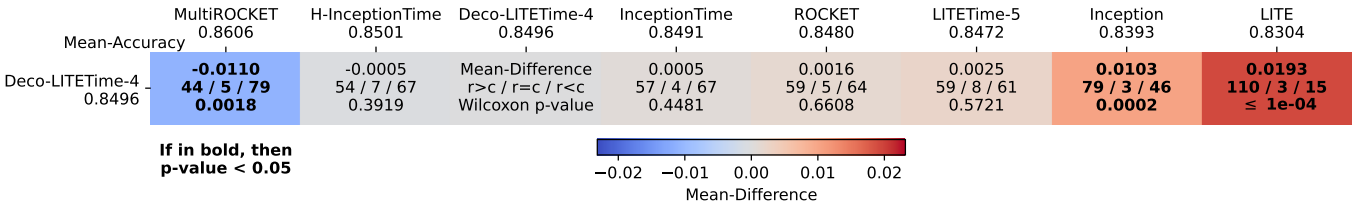

Fig. 5. The Multi-Comparison Matrix applied to show the performance of Deco-LITETime-4 compared to state-of-the-art approaches.

and their decorrelated counterparts. Figure 6 presents a one-vs-one FID score comparison across 128 datasets from the UCR archive. In this figure, the x-axis represents the FID scores computed between a reference model and a base model while the y-axis represents the FID scores computed between the same reference model and corresponding decorrelated model. The results indicate that the decorrelated model produces higher FID scores in 90 datasets, often with a larger margin while the base model results in higher FID scores in 32 datasets. In 6 datasets, there is no observed difference between the two configurations. The p-value computed between these two sets of FID scores is 0.0, confirming that the observed difference is statistically significant. These findings demonstrate that decorrelated learning explicitly enhances feature diversity, encouraging models to learn more distinct representations. By reducing feature redundancy, this approach contributes to better generalization and improved ensemble performance in time series classification tasks.

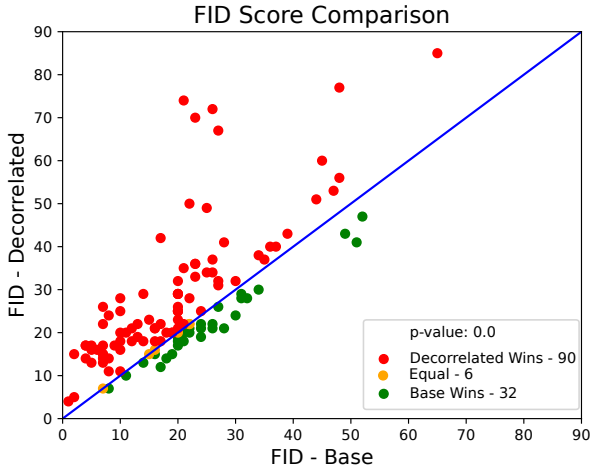

Fig. 6. FID score comparison between two base models and a base-decorrelated model pair.

E. Qualitative Diversity Analysis

To further emphasize our results, we analyze the filter space of base and decorrelated models. The primary objective of the decorrelated loss is to explicitly enhance feature diversity which indirectly drives convolutional filters to learn more distinct representations. To investigate this effect we visualize the learned convolutional filters from both base and decorrelated models to assess filter diversity. Following the default configuration outlined in the LITE model [9], the number of filters in the final layer is 32, resulting in a convolutional filter dimension of 32 × 20 for each base and decorrelated model. To quantitatively analyze filter diversity we employ Dynamic Time Warping (DTW) [39] to measure the similarity between all pairs of filters. Additionally, we employ t-distributed Stochastic Neighbor Embedding (t-SNE) [40] to project the high-dimensional filter representations into a two-dimensional space, enabling straightforward visualization within a Cartesian coordinate system. As illustrated in Figure 7, the convolutional filters from all five base models exhibit a highly similar distribution, forming two distinct clusters

where each cluster contains filters from all base models. This redundancy in the filter space clearly indicates that, despite different initializations, base models tend to learn highly similar filters.

In addition, we analyze the filter space of decorrelated models as shown in Figure 8. This figure includes convolutional filters from the first base model (used as a reference model) along with all four decorrelated models. The diversity between base and decorrelated filters is evident and the decorrelated models themselves also exhibit diversity among them. As previously mentioned, the decorrelation loss for the n-th model is computed as the sum of individual diversity losses with each previously trained model $(1, ..., n-1)$. From the results it seems that the diversity loss between a given decorrelated model and the base model is easier to optimize than that of previously trained decorrelated models. We believe that fine-tuning the weighting of individual diversity losses within the total diversity loss can lead to better results as the optimal trade-off may vary across datasets. From Figure 8, we can observe that the filters of Deco_5 are distributed similarly to those of Deco_2, suggesting that the fifth decorrelated model fails to introduce additional diversity. This can be explained by the fact that most diverse and useful features have already been captured by previously trained decorrelated models, leaving Deco_5 with limited room to learn new and distinctive patterns. Furthermore, it is also worth noting that the 2-model decorrelated ensemble (Deco-LITETime-2) achieves approximately 3.6% higher test accuracy than LITETime-5 which is an ensemble of five base models. These results highlight the effectiveness of our decorrelated learning framework, demonstrating that smaller but more diverse ensembles can outperform larger ensembles composed of redundant models.

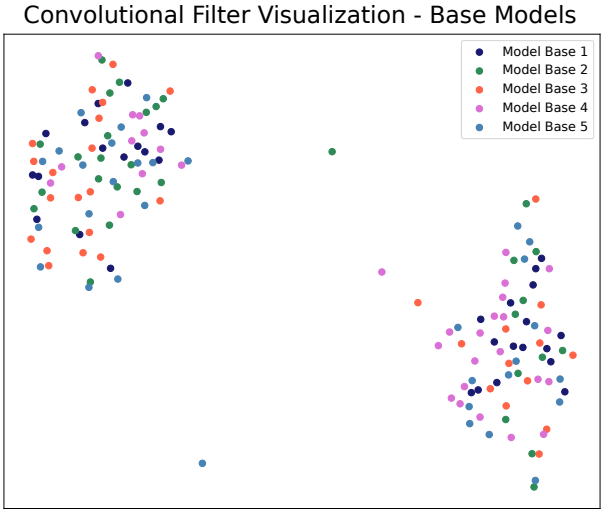

Fig. 7. t-SNE visualization of learned convolutional filters by highlighting redundancy in the filters of base models.

V. Conclusion and Future Work

In this work, we proposed a diversity-driven ensemble learning framework for time series classification that explicitly encourages feature diversity through feature orthogonality loss. By enforcing decorrelation at the feature representation level, our method mitigates redundancy among ensemble members and improves generalization without requiring additional model complexity. Comprehensive experiments on 128 datasets from the UCR archive demonstrated that our approach achieves SOTA performance with fewer models, highlighting the efficiency of diversity-driven ensembling. Both quantitative and qualitative analyses confirmed that enforcing feature diversity results in more complementary feature representations, leading to improved classification accuracy. Future work will explore refining the balance between classification loss and diversity loss, optimizing the trade-off dynamically across different datasets. Additionally, we aim to extend the framework to multivariate time series classification and investigate alternative diversity-promoting strategies beyond feature orthogonality. Further research will also focus on improving computational efficiency by potentially leveraging parallel training strategies to enhance scalability.

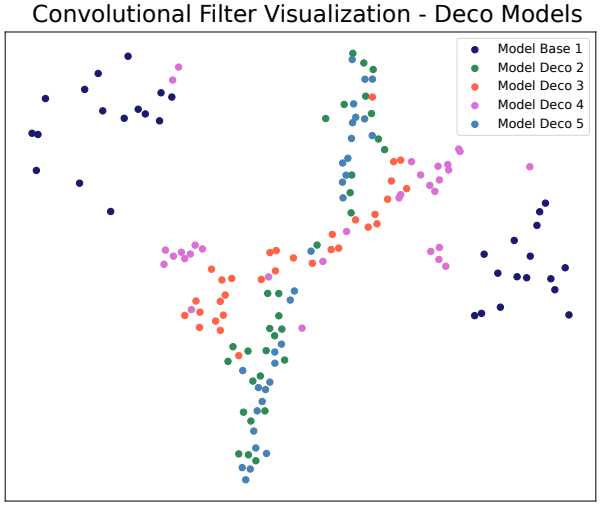

Fig. 8. t-SNE visualization of learned convolutional filters by highlighting diversity between the filters of base and decorrelated models.

Acknowledgment

This work was supported by the ANR DELEGATION project (grant ANR-21-CE23-0014) of the French Agence Nationale de la Recherche. The authors would like to acknowledge the High Performance Computing Center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data. The authors would also like to thank the creators and providers of the UCR Archive.

References

[1] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ digital medicine*, vol. 1, no. 1, p. 18, 2018.

[2] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Systems with Applications*, vol. 105, pp. 233–261, 2018.

[3] F. Yi, Z. Yu, F. Zhuang, X. Zhang, and H. Xiong, "An integrated model for crime prediction using temporal and spatial factors," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1386–1391.

[4] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sensing*, vol. 11, no. 5, p. 523, 2019.

[5] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, "The ucr time series archive," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.

[6] M. Middlehurst, P. Schäfer, and A. Bagnall, "Bake off redux: a review and experimental evaluation of recent time series classification algorithms," *Data Mining and Knowledge Discovery*, pp. 1–74, 2024.

[7] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[8] A. Ismail-Fawaz, M. Devanne, J. Weber, and G. Forestier, "Deep learning for time series classification using new hand-crafted convolution filters," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 972–981.

[9] A. Ismail-Fawaz, M. Devanne, S. Berretti, J. Weber, and G. Forestier, "Lite: Light inception with boosting techniques for time series classification," in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2023, pp. 1–10.

[10] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[11] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall, "Hive-cote 2.0: a new meta ensemble for time series classification," *Machine Learning*, vol. 110, no. 11, pp. 3211–3243, 2021.

[12] E. Ay, M. Devanne, J. Weber, and G. Forestier, "A study of knowledge distillation in fully convolutional network for time series classification," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN55064.2022.9892915

[13] N. Mohammadi Foumani, L. Miller, C. W. Tan, G. I. Webb, G. Forestier, and M. Salehi, "Deep learning for time series classification and extrinsic regression: A current survey," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–45, 2024.

[14] H. A. Dau, D. F. Silva, F. Petitjean, G. Forestier, A. Bagnall, A. Mueen, and E. Keogh, "Optimizing dynamic time warping's window width for time series data mining applications," *Data mining and knowledge discovery*, vol. 32, pp. 1074–1120, 2018.

[15] B. Lucas, A. Shifaz, C. Pelletier, L. O'Neill, N. Zaidi, B. Goethals, F. Petitjean, and G. I. Webb, "Proximity forest: an effective and scalable distance-based classifier for time series," *Data Mining and Knowledge Discovery*, vol. 33, no. 3, pp. 607–635, 2019.

[16] M. A. Bagheri, Q. Gao, and S. Escalera, "Support vector machines with time series distance kernels for action classification," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–7.

[17] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data mining and knowledge discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[18] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 1578–1585.

[19] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[20] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[21] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

[22] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[23] O. Badi, M. Devanne, A. Ismail-Fawaz, J. Abdullayev, V. Lemaire, S. Berretti, J. Weber, and G. Forestier, "Cocalite: A hybrid model combining catch22 and lite for time series classification," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 1229–1236.

[24] H. Lee, K. Lee, D. Hwang, H. Lee, B. Lee, and J. Choo, "On the importance of feature decorrelation for unsupervised representation learning in reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 18 988–19 009.

[25] J. Lines and A. Bagnall, "Time series classification with ensembles of elastic distance measures," *Data Mining and Knowledge Discovery*, vol. 29, pp. 565–592, 2015.

[26] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connection science*, vol. 8, no. 3-4, pp. 385–404, 1996.

[27] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, pp. 181–207, 2003.

[28] S. Yang, W. Deng, M. Wang, J. Du, and J. Hu, "Orthogonality loss: Learning discriminative representations for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2301–2314, 2020.

[29] B. O. Ayinde, T. Inanc, and J. M. Zurada, "Regularizing deep neural networks by enhancing diversity in feature extraction," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2650–2661, 2019.

[30] L. Zhang, Z. Shi, M.-M. Cheng, Y. Liu, J.-W. Bian, J. T. Zhou, G. Zheng, and Z. Zeng, "Nonlinear regression via deep negative correlation learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 982–998, 2019.

[31] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 505–11 515.

[32] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[33] A. Ismail-Fawaz, A. Dempster, C. W. Tan, M. Herrmann, L. Miller, D. F. Schmidt, S. Berretti, J. Weber, M. Devanne, G. Forestier *et al.*, "An approach to multiple comparison benchmark evaluations that is stable under manipulation of the comparate set," *arXiv preprint arXiv:2305.11921*, 2023.

[34] A. Benavoli, G. Corani, and F. Mangili, "Should we really use post-hoc tests based on mean-ranks?" *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 152–161, 2016.

[35] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 196–202.

[36] C. W. Tan, A. Dempster, C. Bergmeir, and G. I. Webb, "Multirocket: multiple pooling operators and transformations for fast and effective time series classification," *Data Mining and Knowledge Discovery*, vol. 36, no. 5, pp. 1623–1646, 2022.

[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[38] A. Ismail-Fawaz, M. Devanne, S. Berretti, J. Weber, and G. Forestier, "Establishing a unified evaluation framework for human motion generation: A comparative analysis of metrics," *Computer Vision and Image Understanding*, vol. 254, p. 104337, 2025.

[39] F. Petitjean, G. Forestier, G. I. Webb, A. E. Nicholson, Y. Chen, and E. Keogh, "Dynamic time warping averaging of time series allows faster and more accurate classification," in *2014 IEEE international conference on data mining*. IEEE, 2014, pp. 470–479.

[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.