Alleviating Overconfidence in Long-Tailed Time Series Classification with Hierarchical Reverse Distillation

Baixin Li
Software college
Northeastern University
Shenyang, China
libaixin@mails.neu.edu.cn

Peibo Duan*
Software college
Northeastern University
Shenyang, China
duanpeibo@swc.neu.edu.cn

Zhipeng Liu
Software college
Northeastern University
Shenyang, China
2310543@stu.neu.edu.cn

Jialu Xu
Software college
Northeastern University
Shenyang, China
xujialu@mails.neu.edu.cn

Fan Zhang
The University of Tokyo
Tokyo, Japan
zhang-fan@g.ecc.u-tokyo.ac.jp

Tianlun Wang
Software college
Northeastern University
Shenyang, China
wangtl2@mails.neu.edu.cn

Bin Zhang
Software college
Northeastern University
Shenyang, China
zhangbin@mail.neu.edu.cn

***Abstract*—Time series classification is a fundamental task with wide-ranging applications. However, existing studies typically assume balanced datasets, which are idealized and overlook the long-tailed nature of real-world data. A key challenge in long-tailed time series data is model overconfidence, with models performing well on head classes but poorly on minority tail classes. Existing approaches attempt to mitigate this issue by enhancing tail class performance, which often comes at the expense of head class accuracy. To address this limitation, we propose a novel Hierarchical Reverse Distillation framework for long-tailed time series classification, named HRD-LT. Specifically, HRD-LT integrates a frequency-based hierarchical data augmentation strategy, generating both weakly and strongly augmented variant series, which enhances the representation of tail samples and mitigates overfitting by generating increasingly challenging samples. Furthermore, we propose a dual-expert collaborative learning strategy with reverse distillation, where the two experts take turns as teacher and student, enabling the overconfident student model to be effectively guided. Extensive experiments on three public datasets, each constructed with versions of varying imbalance ratios, demonstrate that our HRD-LT outperforms state-of-the-art baselines. Compared to the best-performing baselines, HRD-LT achieves improvements of up to 10.36% in terms of accuracy.**

***Index Terms*—Time series classification, Long-tail recognition, Knowledge distillation, Data augmentation**

I. Introduction

Time Series Classification (TSC) is of paramount importance in machine learning and has been widely applied in various domains such as healthcare, transportation management, and quantitative finance [1], [2]. Recent advances in deep learning have led to significant progress in TSC, with existing literature typically developing various backbone architectures and specialized enhancement techniques to capture complex temporal dependencies. However, despite their success in controlled settings, the performance of many methods in real-world applications remains unsatisfactory. A key issue is that most existing approaches are developed and evaluated on idealized balanced datasets, which often fail to reflect the complexities of real-world data. In practice, data typically follow a long-tailed distribution: a small number of dominant classes (*head classes*) account for the majority of samples, while the remaining classes (*tail classes*) are severely underrepresented [1]. For example, in healthcare, rare but critical diseases occur infrequently, while in transportation, traffic accidents are similarly uncommon. This inherent class imbalance poses a significant challenge, often leading to insufficient learning of tail classes. Consequently, models tend to overfit to head classes while under-representing tail classes, a phenomenon we refer to as model *overconfidence*.

To mitigate model overconfidence, researchers in the visual and NLP domains have developed various rebalancing strategies. For instance, contrastive learning-based methods achieve head-tail balance by expanding the feature space of tail classes [3]. Logit adjustment approaches attain head-tail balance by modifying classifier weights [4]. Knowledge distillation methods train a unified model by learning from multiple models trained on imbalanced subsets [5], [6]. Post hoc calibration methods, such as label smoothing and adjusting the classifier's temperature, are used to mitigate model overconfidence [7]. Nevertheless, there is relatively limited research specifically addressing the class imbalance problem in time series data. Several preliminary work [8], [9] has drawn inspiration from other fields and employed contrastive learning-based methods to achieve a more balanced feature space. However, contrastive learning relies heavily on a large number of positive and negative samples, making it less effective in scenarios with extreme

* Corresponding author.

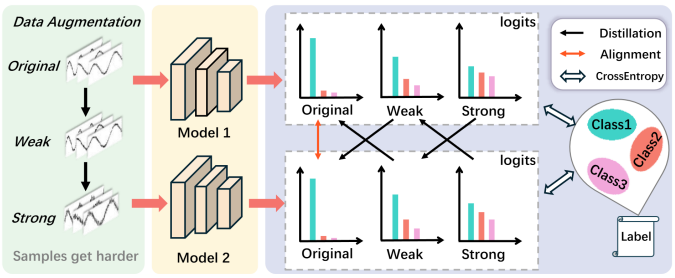


Fig. 1: The overall architecture of HRD-LT. The number of classes is assumed to be three for the sake of illustration.

class imbalance. Although re-sampling [10] can mitigate class imbalance by either replicating tail-class samples or removing head-class samples, it often leads to overfitting or underfitting. ***Overall, existing methods typically improve the learning of tail classes by assigning them larger weights, but this re-weighting strategy typically comes at the expense of head-class performance.***

In this study, we propose a novel framework for long-tailed time series classification, named HRD-LT. Specifically, we first introduce a frequency-based hierarchical time series data augmentation module, which generates both weakly and strongly augmented time series variants that become progressively challenging for prediction, thereby confusing the model to suppress overconfidence. Subsequently, we propose a dual-expert collaborative learning strategy with reverse distillation in which the two expert models learn from each other. Specifically, The logits of the less confident expert are treated as fixed soft labels, guiding the overconfident expert through knowledge distillation, while both models iteratively refine their predictions. Compared to current long-tailed recognition methods, HRD-LT enhances tail performance without relying on class re-weighting, thereby avoiding bias shift and preserving head-class performance. The contributions of this paper can be summarized as: ❶ We introduce a hierarchical data augmentation approach for long-tailed time series classification that progressively generates harder-to-predict samples to suppress model overconfidence. ❷ We propose a collaborative learning framework with reverse knowledge distillation, in which the two experts iteratively serve as teachers for each other, with the less confident model supervising the overconfident model at each iteration. ❸ We conduct extensive experiments on three long-tailed time series datasets derived from three public datasets, with imbalance ratios of 10, 50 and 100. The results show that HRD-LT outperforms its counterparts by up to 10.36% in terms of accuracy.

II. Related Work

A. Time Series Analysis

Deep learning has achieved remarkable success in Time Series Classification (TSC) by automatically learning representations from raw data [11]–[17]. Existing literature typically employs specialized backbone architectures to capture diverse temporal characteristics. Multi-Layer Perceptrons (MLPs) [18]–[20] model time series by stacking fully connected layers to learn global feature representations. Convolutional Neural Networks (CNNs) [21] utilize sliding kernels to extract local discriminative features, establishing a robust baseline for various TSC tasks. Recurrent Neural Networks (RNNs), such as LSTMs and GRUs [22], rely on recurrent connections to model sequential dependencies and temporal evolution effectively. To explicitly capture inter-variable dependencies in multivariate time series, Graph Neural Networks (GNNs) [23]–[26] construct graph structures where nodes represent variables and edges represent their relationships. More recently, Transformer-based models [27]–[31] have achieved state-of-the-art performance by leveraging self-attention mechanisms to capture long-range global dependencies, addressing the limitations of RNNs in processing long sequences. Despite their effectiveness, these models are generally designed for balanced datasets and may struggle to generalize under severe class imbalance.

B. Long-Tailed Recognition

Real-world data often exhibits a long-tailed distribution, where a few head classes dominate while many tail classes are underrepresented. Existing solutions primarily fall into two categories: *class rebalancing* and *information augmentation*. Class rebalancing methods aim to adjust the training distribution, either by resampling data or by reweighting loss functions to assign higher penalties to tail class misclassifications [8], [32]. Information augmentation strategies focus on enriching the representation of tail classes through transfer learning or generative models. However, most existing long-tailed learning methods are tailored for computer vision tasks [33], [34]. Directly applying these methods to time series is challenging due to the complex temporal dynamics and the difficulty in synthesizing realistic time series data without destroying temporal integrity. Therefore, developing specialized strategies for long-tailed TSC remains an open and critical research direction [35], [36].

C. Knowledge Distillation

Knowledge Distillation (KD) is a paradigm originally proposed to transfer knowledge from a cumbersome teacher model to a compact student model [37]. The core idea is to guide the student's learning using soft targets (logits) or intermediate feature representations from the teacher, which contain richer structural information than one-hot labels [38], [39]. Recently, KD has been extended to address the long-tailed problem. In this context, KD is often employed to decouple representation learning from classifier learning, or to transfer the robust feature representations learned from head classes to tail classes [40] . Unlike traditional unidirectional distillation, collaborative learning strategies have also emerged, in which peer models learn from each other [41]–[43]. Our work builds upon these insights but introduces a novel hierarchical reverse distillation mechanism specifically designed to mitigate model overconfidence in time series classification.

III. METHODOLOGY

A. Problem Statement

Given a multi-label time series dataset $\mathcal{D} = \{\mathbf{X}_i, \mathbf{y}_i\}_{i=1}^{N}$ with N samples and C classes, where $\mathbf{X}_i \in \mathbb{R}^{T\times D}$ is the i-th time series sample and $\mathbf{y}_i \in \{0,1\}^C$ is the corresponding label annotation. T and D indicate time steps and the number of feature channels, respectively. We denote $N_{c\in C}$ as the number of samples in the c-th class. Conventional TSC models are typically developed under the assumption of balanced class distributions. However, real-world data often follow a long-tailed distribution, in which class indices are ordered in descending order such that $i < j \Rightarrow N_i \geq N_j$, with $N_1 \gg N_C$. We define the ***imbalance ratio*** as $\rho = N_1/N_C$, Moreover, the top half of the classes are defined as ***head classes***, while the remaining ones are ***tail classes*** [44]. Formally, the prediction can be formulated as follows:

$$\widehat{\mathbf{y}}_i = \mathcal{F}_\phi(\mathbf{X}_i; \Theta_\phi), \tag{1}$$

where $\mathcal{F}_\phi(\cdot)$ denotes the proposed HRD-LT.

B. Overall Framework

The general framework of HRD-LT is shown in Fig.1. First, the hierarchical data augmentation module applies weak and strong augmentations to the original series, producing variants with increased uncertainty. Subsequently, we employ a dual-expert learning strategy, in which two expert models learn collaboratively. Specifically, a reverse distillation loss is designed to enable the less confident model to guide the overconfident one, in which the logits produced by the less confident model serve as soft labels to regularize the outputs of the overconfident model. Finally, an alignment loss is designed to ensure that both models maintain consistent confidence in the original time series. Through this design, HRD-LT alleviates long-tail bias without relying on re-weighting schemes that explicitly favor tail classes at the expense of head-class performance.

C. Hierarchical Data Augmentation

We apply frequency augmentations to the original time series, generating two augmented variant series that includes perturbations at different levels while preserving temporal structures [45]. As the magnitude of the augmentation increases, the model becomes increasingly confused, which helps prevent overconfidence. Moreover, by generating augmented variant series, the model can learn more temporal patterns from tail-class samples.

Specifically, we first employ the Discrete Wavelet Transform (DWT) to decompose the original time series $\mathbf{X}_i$ [1] into detail and approximation coefficients, which capture fine-grained fluctuations and coarse-grained trends. Subsequently, to generate time series variants with different augmentation levels, we add Gaussian noise to the detail and approximation coefficients, respectively. Finally, the augmented coefficients are transformed back to the temporal domain via inverse DWT (iDWT). Formally, the process can be formulated as:

$$\begin{gathered}\mathbf{C}_D, \mathbf{C}_A = \mathrm{DWT}(\mathbf{X}),\\ \mathbf{C}_D^w = \mathbf{C}_D + \epsilon,\ \mathbf{C}_A^s = \mathbf{C}_A + \epsilon,\\ \mathbf{X}^w = \mathrm{iDWT}(\mathbf{C}_A, \mathbf{C}_D^w),\ \mathbf{X}^s = \mathrm{iDWT}(\mathbf{C}_A^s, \mathbf{C}_D).\end{gathered} \tag{2}$$

D. Collaborative Learning with Reverse Distillation

Collaborative learning is a paradigm in which multiple experts work together to mutually enhance and optimize their performance [46], [47]. In this study, we employ two peer experts with identical architectures but different parameters as mutual teachers. The expert exhibiting lower confidence instructs the overconfident counterpart, aiming to alleviate excessive confidence and bolster robustness. This procedure unfolds in three phases. First, the original and augmented samples are separately fed into the two experts to capture temporal dependencies and generate their respective predictive logits,

$$\begin{aligned}\mathbf{z}_1, \mathbf{z}_1^w, \mathbf{z}_1^s &= \mathcal{F}_{\phi_1}(\mathbf{X}), \mathcal{F}_{\phi_1}(\mathbf{X}^w), \mathcal{F}_{\phi_1}(\mathbf{X}^s),\\ \mathbf{z}_2, \mathbf{z}_2^w, \mathbf{z}_2^s &= \mathcal{F}_{\phi_2}(\mathbf{X}), \mathcal{F}_{\phi_2}(\mathbf{X}^w), \mathcal{F}_{\phi_2}(\mathbf{X}^s).\end{aligned} \tag{3}$$

Subsequently, given that time series models are prone to overconfidence specifically when processing long-tailed data, we introduce a reverse distillation method. Here, the expert with lower confidence offers guidance to the overconfident one. In detail, the parameters of the less confident expert are kept fixed, while its logits are employed as soft targets to transfer knowledge to the overconfident expert. Simultaneously, both models collaborate to refine their predictions. The supervision signal for the first expert is defined as:

$$\mathcal{L}_{RD}^1 = D_{KL}(\mathbf{z}_2^w, \mathbf{z}_1) + D_{KL}(\mathbf{z}_2^s, \mathbf{z}_1^w). \tag{4}$$

Similar operations are applied to the second expert, where the first expert is kept fixed to provide guidance. It is noted that since logits represent the probability distribution over classes, we use logits for distillation rather than high-level representations of time series. Moreover, we adopt KL-divergence as the optimization criterion, since unlike mean squared error, which is primarily designed for regression and imposes overly strict constraints, KL-divergence offers a more appropriate measure for aligning probability distributions.

Furthermore, to ensure that reverse distillation is conducted under a comparable confidence baseline, we propose an alignment regularization component. This term enforces consistency in confidence between the predictions of the two experts on the original time series. By minimizing the discrepancy in their logit distributions, we narrow the gap in their intrinsic confidence levels, thereby stabilizing the subsequent knowledge transfer process. Formally, the alignment loss is expressed as:

$$\mathcal{L}_{AL} = D_{JS}(\mathbf{z}_1, \mathbf{z}_2). \tag{5}$$

where $D_{JS}(\cdot)$ denotes the Jensen-Shannon divergence, a symmetric measure of the difference between two probability

[1]For simplicity of notation, the subscript i is omitted in the subsequent derivations.

TABLE I: Dataset Statistics.

Dataset	HAR	ISRUC	Epilepsy
#Train	7352	6013	7360
#val	984	858	1840
#Test	1963	1718	2300
Length	128	3000	178
#Variable	9	10	1
#Class	6	5	5

distributions. Thus, during the inference phase, we employ one of the models for prediction. Finally, the overall optimization objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{class} + \lambda_1(\mathcal{L}^1_{RD} + \mathcal{L}^2_{RD}) + \lambda_2\mathcal{L}_{AL}. \tag{6}$$

where $\mathcal{L}_{class} = \mathcal{L}^1_{class} + \mathcal{L}^2_{class}$ is the classification loss of both models, and λ_1 and λ_2 are hyperparameters striking a balance among terms.

IV. Experiment

A. Dataset and Experimental Setting

1) Dataset: To evaluate the robustness of our method across diverse domains, we select three representative time series classification datasets [48], **HAR**, **ISRUC-S3**, **Epilepsy**, as shown in Table I. Following prior work [49], we construct long-tailed training sets with class distributions following an exponential decay (imbalance ratios $\rho \in \{10, 50, 100\}$), as illustrated in Fig.2(a). Note that the validation and test sets remain unchanged.

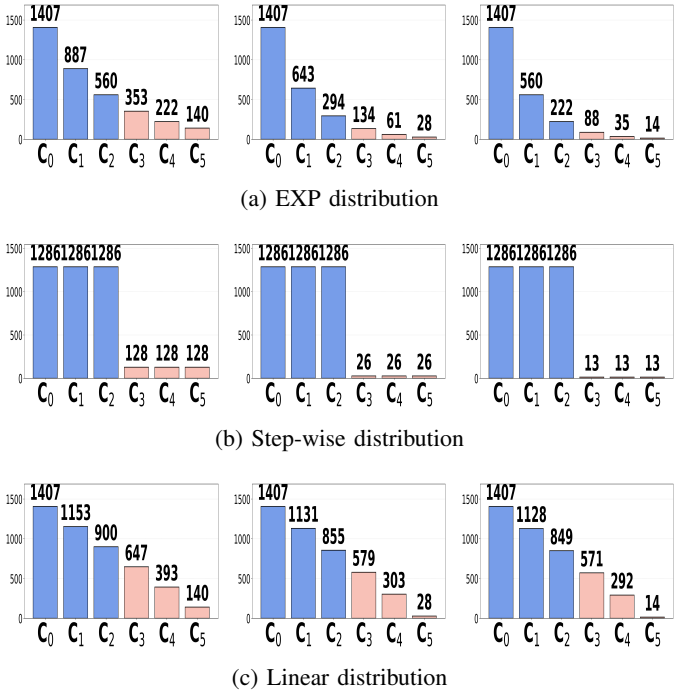
2) Metrics: We employ three evaluation metrics: Accuracy (ACC↑), Macro F1-score (F1↑), and the Matthews Correlation Coefficient (MCC↑), where higher values indicate better performance.

3) Baseline: To validate the effectiveness of our proposed HRD-LT, we compare it against eight standard time-series classification models (DLinear [50], PatchTST [27], iTransformer [51], Timemixer++ [21], and DisMS-TS [48]) and three long-tailed models (Hybrid-PSC [3], IWB [4], and DGMSCL [9]), with DGMSCL being specifically designed for long-tailed time-series classification.

All experiments are conducted on an NVIDIA RTX 4090 24GB GPU. We employed the ADAM optimizer with a initial learning rate of 0.005. For long-tailed models (i.e., Hybrid-PSC, IWB and our HRD-LT), we adopt the temporal encoder from DisMS-TS as a unified backbone. Hyperparameters λ_1 and λ_2 are set to 1. The reported results are averaged over five independent runs for each experiment.

B. Results and Analysis

Tables II and III report the overall and the fine-grained results, respectively. First, we can observe from Table II that, as the imbalance ratio ρ increases, the performance of the models declines. Especially on the Epilepsy dataset with $\rho = 100$, all models exhibit degraded performance. For example, DLinear, iTransformer and TimeMixer++ fail to differentiate between head and tail classes, achieving an MCC of 0. Nonetheless,



(a) EXP distribution

(b) Step-wise distribution

(c) Linear distribution

Fig. 2: Illustration of long-tailed distributions on the HAR-LT dataset. Top: EXP distribution; Middle: Step-wise distribution; Bottom: Linear distribution.

our method consistently achieves the best results, exhibiting statistical significance over all baselines, as confirmed by a t-test conducted at a significance level of $\alpha = 0.01$, which further validates the reliability and reproducibility of our HRD-LT. The superiority can be attributed to our proposed hierarchical data augmentation and reverse distillation strategies, which alleviate model overconfidence.

Furthermore, as shown in Table 3, our method achieves the best performance on both head and tail classes across most datasets. Notably, on the Epilepsy dataset, PatchTST attains the highest accuracy on head classes, as conventional time series models overlook class imbalance and favor head-class predictions. Compared to conventional time series models, long-tailed counterparts show stronger performance on tail classes. However, their performance on head classes is lower than that of our HRD-LT, as their tail-aware strategies emphasize tail classes at the expense of head-class performance.

C. Ablation Study

To validate the contribution of each component in our proposed framework, we conduct ablation studies on the HAR, ISRUC, and Epilepsy datasets with $\rho = 100$. We evaluate four key components, including Weak Augmentation (WA), Strong Augmentation (SA), Reverse Distillation (RD), and the Alignment Regularization term (AR), by removing each respectively.

The results are presented in Table VI, from which we observe that our complete HRD-LT achieves the best performance. Specifically, the removal of the RD module leads to

TABLE II: Overall comparisons on three datasets. **Pink bold** and **Black bold** indicate the best and second-best results, respectively. All experimental results are averaged over five independent runs.

Methods	Metrics	HAR-LT			ISRUC-LT			Epilepsy-LT		
		ρ =10	ρ =50	ρ =100	ρ =10	ρ =50	ρ =100	ρ =10	ρ =50	ρ =100
DLinear	ACC	56.75	50.84	49.61	25.43	24.83	23.72	20.39	20.39	20.39
	F1	53.70	44.82	41.54	20.79	19.56	18.14	6.77	6.77	6.77
	MCC	0.500	0.467	0.458	0.019	0.006	-0.008	0.000	0.000	0.000
PatchTST	ACC	83.95	75.85	74.02	63.32	59.08	53.14	**60.62**	**49.92**	**43.14**
	F1	83.87	74.45	72.47	59.36	49.35	43.75	**57.81**	**46.48**	**39.18**
	MCC	0.809	0.717	0.699	0.544	0.488	0.400	**0.532**	**0.417**	**0.347**
iTransformer	ACC	91.34	80.43	75.29	31.19	30.84	30.84	20.39	20.39	20.39
	F1	91.16	79.39	72.82	13.66	12.33	10.22	6.77	6.77	6.77
	MCC	0.896	0.769	0.712	0.023	0.004	-0.012	0.000	0.000	0.000
TimeMixer++	ACC	90.52	73.61	69.63	56.77	47.48	42.20	20.39	20.39	20.39
	F1	90.37	71.52	65.45	56.77	31.21	24.19	20.39	6.77	6.77
	MCC	0.888	0.691	0.652	0.462	0.329	0.230	0.000	0.000	0.000
DisMS-TS	ACC	**93.36**	**91.46**	**86.47**	80.22	72.69	62.82	52.64	47.02	40.22
	F1	**93.07**	**91.26**	**86.16**	75.08	69.22	57.62	45.17	39.70	32.74
	MCC	**0.924**	**0.904**	**0.846**	0.723	0.665	0.547	0.457	0.416	0.364
Hybrid-PSC	ACC	92.53	91.39	85.79	74.18	69.56	60.18	50.62	44.62	37.78
	F1	91.46	90.45	83.57	71.39	69.22	57.62	44.85	39.70	32.74
	MCC	0.915	0.901	0.832	0.679	0.645	0.539	0.425	0.382	0.324
IWB	ACC	91.88	91.10	86.23	**80.97**	**73.16**	**65.34**	51.74	45.12	37.89
	F1	90.47	88.91	85.95	**77.83**	**69.68**	**62.95**	47.15	37.54	29.38
	MCC	0.902	0.891	0.843	**0.743**	**0.676**	**0.590**	0.443	0.396	0.331
DGMSCL	ACC	92.20	86.04	86.36	77.64	65.19	58.73	43.97	37.39	35.48
	F1	92.06	85.56	85.92	72.59	50.68	42.15	35.80	29.70	27.42
	MCC	0.907	0.836	0.839	0.695	0.578	0.523	0.376	0.294	0.255
HRD-LT	**ACC**	**96.41**	**94.24**	**90.51**	**82.19**	**76.16**	**71.23**	**61.83**	**59.93**	**53.50**
	F1	**96.38**	**94.10**	**90.30**	**78.62**	**72.51**	**65.60**	**58.84**	**55.95**	**45.72**
	MCC	**0.951**	**0.931**	**0.890**	**0.765**	**0.719**	**0.620**	**0.552**	**0.536**	**0.445**

the most significant performance degradation, with the ACC decreasing by 3.22%, 3.64%, and 4.50% on the HAR, ISRUC, and Epilepsy datasets, respectively. Similarly, the exclusion of WA and SA results in ACC drops of up to 1.65% and 1.35% on the HAR dataset. Furthermore, removing the AR term leads to an average ACC decrease of approximately 1.26% across the three datasets. These observations indicate that all components positively contribute to the prediction and further demonstrate the effectiveness of HRD-LT.

D. Visualization Analysis

To intuitively assess the quality of representations learned by our method, we visualize the embeddings obtained from three representative methods, DGMSCL, PatchTST and TimeMixer++, alongside our approach on the HAR dataset with an imbalance ratio of 100, as shown in Fig. 3. The embeddings are projected into a two-dimensional space via t-SNE. Compared to DGMSCL, PatchTST and TimeMixer++, our HRD-LT demonstrates clearer boundaries between different classes.

E. Hyperparameter Sensitivity Analysis

We evaluate the sensitivity of HRD-LT to the reverse distillation coefficient λ_1 and alignment weight λ_2 in Eq. (6). All experiments in this section are conducted on the HAR dataset under the imbalance ratio ($\rho = 100$). Specifically, we adopt a grid search strategy to identify the optimal hyperparameter settings within the set $\{0.1, 0.5, 1, 5, 10\}$. As shown in Fig. 4, optimal performance for both coefficients is achieved at 1.0. A moderate λ_1 provides effective guidance to alleviate model overconfidence, whereas extreme values result in insufficient knowledge transfer or suppressed classification signals. For λ_2, the performance remains stable across various settings, indicating that aligning logit distributions effectively maintains consistency between experts. Overall, HRD-LT exhibits high robustness to hyperparameter variations under severe data imbalance, confirming the stability of our collaborative learning framework.

F. Impact of Loss Functions

To validate the effectiveness of our distillation strategy, we compare the proposed KL-divergence loss with the standard Mean Squared Error (MSE) loss across three datasets under imbalance ratios of $\rho \in \{50, 100\}$. The results in Table VII demonstrate that the KL-based objective consistently yields superior performance. Taking the HAR dataset ($\rho = 100$) as an example, the KL loss achieves an accuracy of 90.51%, surpassing the MSE loss (88.73%) by a clear margin. Similar performance gains are evident on the ISRUC and Epilepsy datasets. This advantage stems from the flexible distribution alignment inherent in KL-divergence. Unlike the rigid element-wise logit matching of MSE, this soft probability matching effectively captures inter-class correlations, which helps stabilize minority class representations and mitigates majority bias under extreme imbalance.

TABLE III: Detailed accuracy (ACC %) comparison.

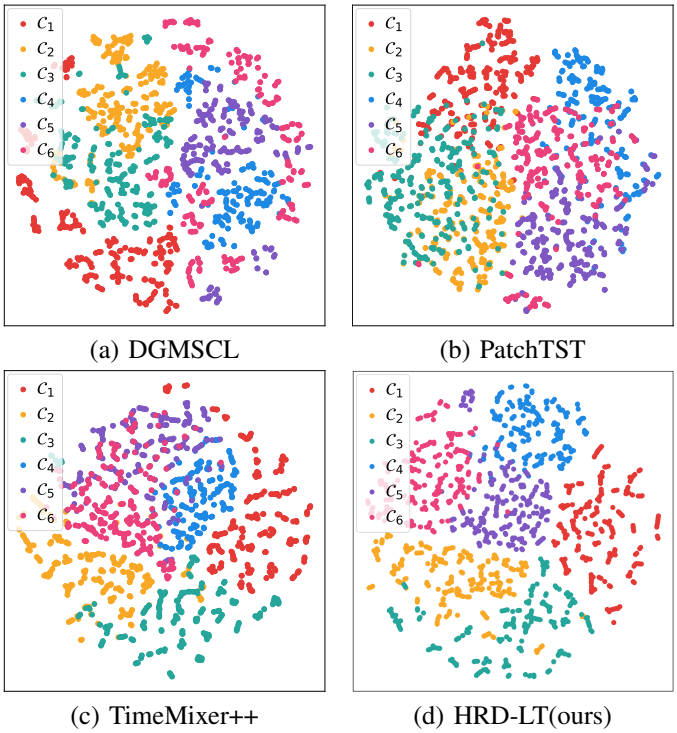
Methods	HAR-10		ISRUC-10		Epilepsy-10	
	Head	Tail	Head	Tail	Head	Tail
DLinear	90.08	10.53	37.71	11.34	50.00	0.00
PatchTST	78.50	86.40	82.50	44.15	**73.40**	45.85
iTransformer	91.50	90.10	51.50	5.20	50.00	0.00
TimeMixer++	88.40	85.30	78.20	25.40	50.00	0.00
DisMS-TS	**95.50**	89.20	88.10	58.40	52.80	48.50
Hybrid-PSC	94.20	89.80	83.50	52.60	53.40	42.10
IWB	94.80	**89.90**	**87.80**	55.30	56.50	41.20
DGMSCL	89.50	86.40	87.20	48.50	32.40	**46.80**
HRD-LT	**97.10**	**95.80**	**89.50**	**69.80**	**68.50**	**57.20**
Methods	HAR-50		ISRUC-50		Epilepsy-50	
	Head	Tail	Head	Tail	Head	Tail
DLinear	89.72	8.85	36.25	10.95	50.00	0.00
PatchTST	73.35	77.90	80.70	34.62	**70.15**	35.75
iTransformer	90.05	69.04	50.65	0.00	50.00	0.00
TimeMixer++	86.26	52.17	74.67	13.8	50.00	0.00
DisMS-TS	**94.13**	84.66	84.65	43.32	49.32	41.25
Hybrid-PSC	93.01	86.69	80.54	41.12	49.61	36.24
IWB	93.83	**88.24**	**85.49**	**45.72**	52.46	36.73
DGMSCL	87.18	81.85	85.47	35.65	26.99	**43.55**
HRD-LT	**95.72**	**93.15**	**86.86**	**63.57**	**65.97**	**52.36**
Methods	HAR-100		ISRUC-100		Epilepsy-100	
	Head	Tail	Head	Tail	Head	Tail
DLinear	87.23	7.90	34.46	9.25	50.00	0.00
PatchTST	72.37	74.99	78.40	27.90	**66.91**	26.59
iTransformer	90.34	59.04	49.52	0.00	50.00	0.00
TimeMixer++	86.26	52.17	69.22	6.86	50.00	0.00
DisMS-TS	**91.17**	80.97	80.72	40.01	46.02	37.44
Hybrid-PSC	88.95	81.32	77.25	36.84	44.17	32.91
IWB	90.30	82.76	82.15	**42.67**	43.58	33.26
DGMSCL	88.18	**82.85**	**82.22**	32.38	18.81	**45.90**
HRD-LT	**93.70**	**86.99**	**83.90**	**61.78**	**61.06**	**47.30**

G. Robustness to Different Imbalance Distributions

To comprehensively evaluate the robustness of our method under varying distribution patterns, we introduce two additional imbalance configurations: **Step-wise** and **Linear**. In these settings, the number of samples per class exhibits a step-wise descent and a linear decay, respectively. We conduct experiments on the representative HAR dataset using three imbalance ratios: $\rho \in \{10, 50, 100\}$. The detailed data distributions for these settings are illustrated in Fig. 2(b) and 2(c), respectively. The corresponding comparative experimental results are presented in Table IV and Table V, which further demonstrate that our method significantly outperforms competing baselines across diverse distribution profiles.

V. Conclusion

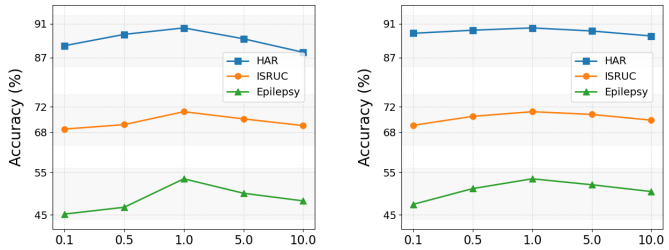
In this paper, we revisit long-tailed time series classification, as existing approaches often enhance tail-class representation at the expense of head-class performance to mitigate model overconfidence. To overcome this limitation, we proposed HRD-LT, a hierarchical reverse distillation framework that regulates model confidence through frequency-aware data augmentation and dual-expert collaborative learning. By enabling experts to alternately guide each other via reverse distillation, HRD-LT enhances tail-class representations while preserving the discriminability of head classes. Extensive experiments on multiple benchmarks with diverse imbalance patterns consistently demonstrate the effectiveness and robustness of the proposed method.



(a) DGMSCL (b) PatchTST

(c) TimeMixer++ (d) HRD-LT(ours)

Fig. 3: t-SNE visualization analysis.



(a) Effect of λ_1 (b) Effect of λ_2

Fig. 4: Hyperparameter sensitivity analysis on the HAR dataset with an imbalance ratio $\rho = 100$. The horizontal axis represents the values of λ_1 and λ_2 respectively, while the vertical axis denotes the classification accuracy (ACC %).

References

[1] Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binquan Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama, "A systematic review on long-tailed learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[2] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng, "Deep long-tailed learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10795–10816, 2023.

TABLE IV: Overall performance and head/tail accuracy on the HAR-LT dataset under the **Step-wise** long-tailed distribution.

Methods	ρ =10					ρ =50					ρ =100				
	ACC	F1	MCC	Head	Tail	ACC	F1	MCC	Head	Tail	ACC	F1	MCC	Head	Tail
DLinear	55.20	52.85	0.520	89.45	9.74	50.15	43.80	0.455	87.25	8.20	48.90	40.30	0.446	86.15	6.95
PatchTST	79.97	79.50	0.760	80.64	78.55	73.96	72.94	0.689	80.72	65.51	65.43	62.95	0.591	81.61	46.41
iTransformer	89.82	88.67	0.872	91.76	88.54	79.54	78.19	0.769	86.15	64.92	73.26	71.54	0.688	86.26	60.15
TimeMixer++	88.20	87.91	0.859	92.52	83.41	71.93	68.00	0.684	87.81	54.00	66.07	62.46	0.603	89.76	38.38
DisMS-TS	**91.78**	**90.84**	**0.908**	**95.42**	85.73	**90.18**	**89.45**	**0.897**	**93.68**	84.67	**86.47**	**86.16**	**0.846**	**91.54**	**80.65**
Hybrid-PSC	90.27	89.76	0.891	93.35	86.91	89.42	87.82	0.883	91.73	85.54	83.57	80.94	0.793	87.49	77.83
IWB	89.34	88.52	0.887	91.14	**87.38**	88.95	87.24	0.879	91.15	**85.38**	85.95	84.43	0.833	89.87	78.56
DGMSCL	91.13	90.75	0.903	94.22	86.61	81.71	81.13	0.781	88.79	73.17	73.40	71.05	0.688	90.37	54.72
HRD-LT (ours)	**95.65**	**95.67**	**0.957**	**96.58**	**94.59**	**93.73**	**93.71**	**0.930**	**94.84**	**92.82**	**89.74**	**89.23**	**0.893**	**93.04**	**85.34**
Improvements	↑ 4.22%*	↑ 5.32%*	↑ 5.40%*	↑ 1.22%*	↑ 6.83%*	↑ 3.94%*	↑ 4.76%*	↑ 3.68%*	↑ 1.24%*	↑ 8.51%*	↑ 3.78%*	↑ 3.56%*	↑ 5.56%*	↑ 1.64%*	↑ 5.82%*

TABLE V: Overall performance and head/tail accuracy on the HAR-LT dataset under the **Linear** long-tailed distribution.

Methods	ρ =10					ρ =50					ρ =100				
	ACC	F1	MCC	Head	Tail	ACC	F1	MCC	Head	Tail	ACC	F1	MCC	Head	Tail
DLinear	56.87	54.32	0.551	91.36	10.43	49.85	43.25	0.452	86.14	7.92	47.45	39.88	0.441	85.79	7.23
PatchTST	86.75	86.75	0.842	92.32	81.43	76.36	74.38	0.724	80.18	72.61	74.70	69.97	0.711	81.43	67.82
iTransformer	88.64	87.43	0.862	90.38	87.92	78.92	77.34	0.758	85.67	63.12	72.83	70.96	0.678	83.15	58.02
TimeMixer++	**93.63**	**93.52**	**0.924**	**96.57**	**90.64**	86.04	85.77	0.834	87.79	**84.46**	82.06	81.25	0.791	84.99	**78.91**
DisMS-TS	91.73	90.52	0.907	94.26	87.15	**90.53**	**89.68**	**0.903**	**93.12**	84.28	**84.52**	**83.87**	**0.831**	89.46	78.39
Hybrid-PSC	89.54	88.72	0.886	92.42	85.67	88.62	87.76	0.875	91.38	85.15	81.28	80.54	0.809	87.54	76.34
IWB	88.28	87.45	0.874	92.15	84.93	86.45	85.63	0.868	91.85	82.58	82.67	81.12	0.815	89.05	78.83
DGMSCL	90.23	90.08	0.884	88.13	92.15	81.59	79.52	0.788	88.61	74.35	78.19	71.86	0.758	**91.45**	64.70
HRD-LT (ours)	**95.89**	**95.86**	**0.952**	**96.62**	**95.01**	**94.44**	**94.21**	**0.933**	**96.34**	**91.82**	**91.19**	**90.60**	**0.894**	**95.42**	**86.47**
Improvements	↑ 2.41%*	↑ 2.50%*	↑ 3.03%*	↑ 0.05%*	↑ 4.82%*	↑ 4.32%*	↑ 5.05%*	↑ 3.32%*	↑ 3.46%*	↑ 7.83%*	↑ 7.89%*	↑ 8.02%*	↑ 7.58%*	↑ 4.34%*	↑ 9.58%*

TABLE VI: Ablation results on three datasets with $\rho = 100$.

Variants	HAR	ISRUC	Epilepsy
w/o WA	88.86/88.27	69.20/64.34	50.30/44.98
w/o SA	89.16/88.89	69.51/64.53	51.47/44.88
w/o RD	87.29/86.98	67.59/63.18	49.00/43.95
w/o AR	89.01/88.53	70.28/64.69	52.17/45.05
HRD-LT	**90.51/90.30**	**71.23/65.60**	**53.50/45.72**

Format: Acc(%)/F1(%)

TABLE VII: Performance comparison of different distillation loss functions across three datasets ($\rho = 50, 100$).

LOSS	Metrics	HAR-LT		ISRUC-LT		Epilepsy-LT	
		ρ =50	ρ =100	ρ =50	ρ =100	ρ =50	ρ =100
MSE	ACC	92.87	88.73	74.28	68.95	57.41	51.22
	F1	92.65	88.48	70.83	63.17	53.28	43.15
	MCC	0.915	0.872	0.695	0.591	0.498	0.407
KL	ACC	**94.24**	**90.51**	**76.16**	**71.23**	**59.93**	**53.50**
	F1	**94.10**	**90.30**	**72.51**	**65.60**	**55.95**	**45.72**
	MCC	**0.931**	**0.890**	**0.719**	**0.620**	**0.536**	**0.445**

[3] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 943–952.

[4] Wenqi Dang, Zhou Yang, Weisheng Dong, Xin Li, and Guangming Shi, "Inverse weight-balancing for deep long-tailed learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 11713–11721.

[5] Liuyu Xiang, Guiguang Ding, and Jungong Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *European conference on computer vision*. Springer, 2020, pp. 247–263.

[6] Harsh Rangwani, Pradipto Mondal, Mayank Mishra, Ashish Ramayee Asokan, and R Venkatesh Babu, "Deit-lt: Distillation strikes back for vision transformer training on long-tailed datasets," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 23396–23406.

[7] Yin-Yin He, Peizhen Zhang, Xiu-Shen Wei, Xiangyu Zhang, and Jian Sun, "Relieving long-tailed instance segmentation via pairwise class balance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7000–7009.

[8] Pengkun Wang, Xu Wang, Binwu Wang, Yudong Zhang, Lei Bai, and Yang Wang, "Long-tailed time series classification via feature space rebalancing," in *International Conference on Database Systems for Advanced Applications*. Springer, 2023, pp. 151–166.

[9] Lipeng Qian, Qiong Zuo, Dahu Li, and Hong Zhu, "Dgmscl: A dynamic graph mixed supervised contrastive learning approach for class imbalanced multivariate time series classification," *Neural Networks*, vol. 185, pp. 107131, 2025.

[10] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[11] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang, "TFB: Towards comprehensive and fair benchmarking of time series forecasting methods," in *Proc. VLDB Endow.*, 2024, pp. 2363–2377.

[12] Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang, "DUET: Dual clustering enhanced multivariate time series forecasting," in *SIGKDD*, 2025, pp. 1185–1196.

[13] Xingang Guo, Yaxin Li, Xiangyi Kong, Yilan Jiang, Xiayu Zhao, Zhihua Gong, Yufan Zhang, Daixuan Li, Tianle Sang, Beixiao Zhu, et al., "Toward engineering agi: Benchmarking the engineering design capabilities of llms," *arXiv preprint arXiv:2509.16204*, 2025.

[14] Xiangfei Qiu, Zhe Li, Wanghui Qiu, Shiyan Hu, Lekui Zhou, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Aoying Zhou, Zhenli Sheng, Jilin Hu, Christian S. Jensen, and Bin Yang, "Tab: Unified benchmarking of time series anomaly detection methods," in *Proc. VLDB Endow.*, 2025, pp. 2775–2789.

[15] Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xvyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang, "Dbloss: Decomposition-based loss function for time series forecasting," in *NeurIPS*, 2025.

[16] Zhipeng Liu, Peibo Duan, Xuan Tang, Haodong Jing, Mingyang Geng, Yongsheng Huang, Jialu Xu, Bin Zhang, and Binwu Wang, "We

need a more robust classifier: Dual causal learning empowers domain-incremental time series classification," *arXiv preprint arXiv:2601.10312*, 2026.

[17] Fan Zhang, Jiabin Luo, Zheng Zhang, Shuanghong Huang, Zhipeng Liu, and Yu Chen, "Beyond visual realism: Toward reliable financial time series generation," *arXiv preprint arXiv:2601.12990*, 2026.

[18] Sijia Peng, Yun Xiong, Yangyong Zhu, and Zhiqiang Shen, "Semantics-aware patch encoding and hierarchical dependency modeling for long-term time series forecasting," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 2269–2280.

[19] Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang, "Rethinking irregular time series forecasting: A simple yet effective baseline," in *AAAI*, 2026.

[20] Maohao Shen, Jacky Y Zhang, Leihao Chen, Weiman Yan, Neel Jani, Brad Sutton, and Oluwasanmi Koyejo, "Labeling cost sensitive batch active learning for brain tumor segmentation," in *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE, 2021, pp. 1269–1273.

[21] Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin, "Timemixer++: A general time series pattern machine for universal predictive analysis," *arXiv preprint arXiv:2410.16032*, 2024.

[22] Yaxuan Kong, Zepu Wang, Yuqi Nie, Tian Zhou, Stefan Zohren, Yuxuan Liang, Peng Sun, and Qingsong Wen, "Unlocking the power of lstm for long term time series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, 2025, vol. 39, pp. 11968–11976.

[23] Binwu Wang, Yudong Zhang, Xu Wang, Pengkun Wang, Zhengyang Zhou, Lei Bai, and Yang Wang, "Pattern expansion and consolidation on evolving graphs for continual traffic prediction," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 2223–2232.

[24] Binwu Wang, Pengkun Wang, Yudong Zhang, Xu Wang, Zhengyang Zhou, Lei Bai, and Yang Wang, "Towards dynamic spatial-temporal graph learning: A decoupled perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 9089–9097.

[25] Zhipeng Liu, Peibo Duan, Xiaosha Xue, Changsheng Zhang, Wenwei Yue, and Bin Zhang, "A dynamic hypergraph attention network: Capturing market-wide spatiotemporal dependencies for stock selection," *Applied Soft Computing*, vol. 169, pp. 112524, 2025.

[26] Zhipeng Liu, Peibo Duan, Qi Chu, Levin Kuhlmann, Changsheng Zhang, Wenwei Yue, Xuan Tang, and Bin Zhang, "An attributed multiplex network enabled gnn-based stock predictor with observable and non-observable information," *Expert Systems with Applications*, p. 129018, 2025.

[27] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[28] Zhipeng Liu, Peibo Duan, Xuan Tang, Baixin Li, Yongsheng Huang, Mingyang Geng, Changsheng Zhang, Bin Zhang, and Binwu Wang, "Timeformer: Transformer with attention modulation empowered by temporal characteristics for time series forecasting," *Expert Systems with Applications*, p. 131040, 2025.

[29] Xiaosha Xue, Peibo Duan, Zhipeng Liu, Qi Chu, Changsheng Zhang, and Bin Zhang, "Gated fusion enhanced multi-scale hierarchical graph convolutional network for stock movement prediction," in *International Conference on Neural Information Processing*. Springer, 2025, pp. 381–395.

[30] Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu, "Dag: A dual correlation network for time series forecasting with exogenous variables," *arXiv preprint arXiv:2509.14933*, 2025.

[31] Weiman Yan, Yi-Chia Chang, and Wanyu Zhao, "Stiff circuit system modeling via transformer," *arXiv preprint arXiv:2510.24727*, 2025.

[32] Binwu Wang, Pengkun Wang, Wei Xu, Xu Wang, Yudong Zhang, Kun Wang, and Yang Wang, "Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning," in *the twelfth international conference on learning representations*, 2024.

[33] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang, "Trustworthy long-tailed classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6970–6979.

[34] Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang, "Invariant feature learning for generalized long-tailed classification," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–726.

[35] Pengkun Wang, Zhe Zhao, HaiBin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang, "Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems," *Advances in Neural Information Processing Systems*, vol. 37, pp. 64915–64941, 2024.

[36] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton Van den Hengel, "Retrieval augmented classification for long-tail visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 6959–6969.

[37] Zhipeng Liu, Peibo Duan, Mingyang Geng, and Bin Zhang, "A distillation-based future-aware graph neural network for stock trend prediction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[38] Laiqiao Qin, Tianqing Zhu, Wanlei Zhou, and Philip S Yu, "Knowledge distillation in federated learning: A survey on long lasting challenges and new solutions," *International Journal of Intelligent Systems*, vol. 2025, no. 1, pp. 7406934, 2025.

[39] Wentao Zhang, Xupeng Miao, Yingxia Shao, Jiawei Jiang, Lei Chen, Olivier Ruas, and Bin Cui, "Reliable data distillation on graph convolutional network," in *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, 2020, pp. 1399–1414.

[40] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao, "Amalgamating knowledge from heterogeneous graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15709–15718.

[41] Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, et al., "Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions," *Artificial Intelligence Review*, vol. 59, no. 1, pp. 17, 2026.

[42] Cheng Yang, Jiawei Liu, and Chuan Shi, "Extract the knowledge of graph neural networks and go beyond it: An effective knowledge distillation framework," in *Proceedings of the web conference 2021*, 2021, pp. 1227–1237.

[43] Weiman Yan, Ernest Wu, and Elyse Rosenbaum, "New loss function for learning dielectric thickness distributions and generative modeling of breakdown lifetime," in *2025 IEEE International Reliability Physics Symposium (IRPS)*. IEEE, 2025, pp. 1–9.

[44] Chih-Hui Ho, Kuan-Chuan Peng, and Nuno Vasconcelos, "Long-tailed anomaly detection with learnable class names," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12435–12446.

[45] Yucheng Wang, Yuecong Xu, Jianfei Yang, Min Wu, Xiaoli Li, Lihua Xie, and Zhenghua Chen, "Graph-aware contrasting for multivariate time-series classification," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, vol. 38, pp. 15725–15734.

[46] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton, "Large scale distributed neural network training through online distillation," *arXiv preprint arXiv:1804.03235*, 2018.

[47] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo, "Online knowledge distillation via collaborative learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11020–11029.

[48] Zhipeng Liu, Peibo Duan, Binwu Wang, Xuan Tang, Qi Chu, Changsheng Zhang, Yongsheng Huang, and Bin Zhang, "Disms-ts: Eliminating redundant multi-scale features for time series classification," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 10817–10826.

[49] Chih-Hui Ho, Kuan-Chuan Peng, and Nuno Vasconcelos, "Long-tailed anomaly detection with learnable class names," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12435–12446.

[50] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu, "Are transformers effective for time series forecasting?," in *Proceedings of the AAAI conference on artificial intelligence*, 2023, vol. 37, pp. 11121–11128.

[51] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.