A Conformal Predictive Measure for Assessing Catastrophic Forgetting

Ioannis Pitsiorlas[†*], Nour Jamoussi[†*], Marios Kountouris[†‡]
[†]Communication Systems Department, EURECOM, France
[‡]Andalusian Institute of Data Science and Computational Intelligence (DaSCI)
Department of Computer Science and Artificial Intelligence, University of Granada, Spain
[*]*Equal contribution*

***Abstract*—This work introduces a novel methodology for assessing catastrophic forgetting (CF) in continual learning. We propose a new conformal prediction (CP)-based metric, termed the *Conformal Prediction Confidence Factor (CPCF)*, to quantify and evaluate CF effectively. Our framework leverages adaptive CP to estimate forgetting by monitoring the model's confidence on previously learned tasks. This approach provides a dynamic and practical solution for monitoring and measuring CF of previous tasks as new ones are introduced, offering greater suitability for real-world applications. Experimental results on four benchmark datasets demonstrate a strong correlation between CPCF and the accuracy of previous tasks, validating the reliability and interpretability of the proposed metric. Our results highlight the potential of CPCF as a robust and effective tool for assessing and understanding CF in dynamic learning environments.**

***Index Terms*—conformal prediction, catastrophic forgetting, continual learning**

I. Introduction

Current machine learning (ML) models are predominantly designed for single-task applications, where they undergo iterative training over multiple epochs on a static dataset until convergence is reached [1]. However, significant challenges arise when models are trained exclusively on fixed datasets, particularly in real-world scenarios where new tasks or classes may frequently emerge over time. For instance, the transition to 6G networks in wireless communications calls for adaptive resource allocation models that can accommodate new devices or services without requiring complete retraining. As user demands and network configurations change, traditional ML models often fail to generalize effectively, rendering full retraining costly and impractical in dynamic environments. A similar issue arises in healthcare, where ML models used for diagnostics are typically trained on known diseases but must also adapt to new conditions, such as emerging virus strains or rare diseases, without compromising accuracy on previously learned cases [1]. These challenges highlight the need for ML models that can continuously learn and adapt to dynamic, real-world environments while preserving their performance on previously learned tasks.

To address this challenge, continual learning (CL), also known as lifelong or incremental learning, has gained considerable attention. CL aims to incrementally learn from a sequence of tasks or data points while preserving the performance of the model as if all data were available concurrently [2]. A key requirement of a CL system is its ability to adapt to new tasks or shifts in data distribution without relying on access to previously encountered data [3]. This capability is particularly critical in scenarios where data privacy is a priority, such as under General Data Protection Regulation (GDPR) compliance, where organizations are obligated to restrict data usage and storage beyond a specified retention period. These restrictions significantly complicate the process of reusing historical datasets for retraining. Moreover, the high computational cost of retraining, especially for large-scale models, presents another major barrier when integrating new data [4]. By addressing both data privacy and computational efficiency, CL offers a promising solution for enabling models to evolve over time while remaining compliant with privacy regulations and minimizing resource demands.

Continual learning systems often suffer from catastrophic forgetting (CF), a phenomenon in which previously acquired knowledge is overwritten as new information is learned [5]. This typically occurs when a model adapts to a shift in the data distribution, resulting in a degradation in performance on earlier tasks or previously encountered distributions.

Distribution shifts, such as covariate, label, and concept shifts, refer to changes in the underlying data distribution encountered by a model during training or deployment, often leading to significant performance degradation. Covariate shift involves changes in the input distribution, while the relationship between the input and output remains stable [6]. Label shift refers to changes in the distribution of output labels while the input distribution stays the same [7]. Concept shift refers to changes in the input-output relationship itself, which often arises when new tasks are introduced [8].

Over the years, numerous methods have been proposed in CL to address the challenge of CF. These approaches generally include regularization-based approaches, such as Elastic Weight Consolidation (EWC) [9], Online Structured Laplace Approximations [10], replay-based approaches, such as Reservoir Sampling [11], and optimization-based strategies, such as Gradient Episodic Memory, among others [12]. A significant limitation of current CL methods is their tendency to overlook predictive uncertainty, often leading to overconfident yet unreliable decisions [13]. This limitation becomes particularly pronounced and problematic in real-world applications, where models encounter dynamically evolving

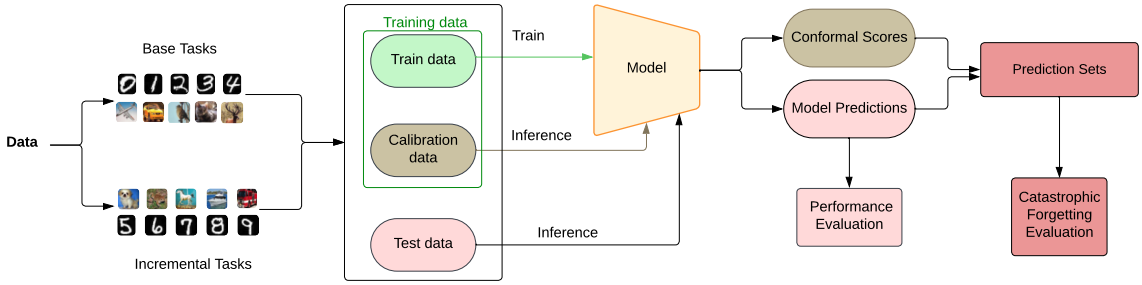

Fig. 1: Proposed framework overview for assessing catastrophic forgetting in continual learning via conformal prediction.

tasks and data distributions. In high-stakes domains such as healthcare (e.g., risk of misdiagnosis), autonomous systems, or intrusion detection, overconfidence in predictions can result in critical errors, compromised safety, and severe real-world consequences [14], [15].

Quantifying predictive confidence is essential not only to improve decision-making, but also to effectively assess and mitigate CF. Traditional methods, such as softmax-based distillation, often produce overconfident predictions and fail to capture the underlying uncertainty in model parameters or data distributions [16]. In contrast, uncertainty-aware approaches, including those leveraging Bayesian techniques or ensemble methods, have shown promising potential. However, these methods are typically computationally intensive and difficult to scale in CL settings, where efficiency and adaptability are critical requirements [16].

Conformal prediction (CP) has been successfully applied in various domains, including medical diagnostics, anomaly detection, and uncertainty quantification, to provide calibrated confidence measures for model predictions [17]. As a model-agnostic and principled framework, CP offers an effective way to quantify uncertainty by providing calibrated confidence levels that remain reliable and can adapt to dynamic learning environments.

In this work, we propose a conformal predictive measure for assessing CF, termed the Conformal Prediction Confidence Factor (CPCF). By integrating CP into the CL pipeline, our approach enables an online evaluation framework for CF, enhancing the evaluation process to better reflect the challenges of dynamic real-world environments. Our empirical results demonstrate a strong correlation between the proposed CPCF and model accuracy, in two different learning settings and four benchmark datasets, underscoring its effectiveness in detecting CF.

II. Proposed Methodology

In this section, we introduce *Adaptive CP* [17] for classification tasks and propose a new metric, the CPCF, to assess CF in CL settings. An overview of the proposed framework is shown in Figure 1.

A. Data Splitting

To implement CP, the training dataset is split into two subsets according to a predefined calibration ratio.

- **Effective Training Data:** The portion of the dataset that is used to train the model.
- **Calibration Data:** The remaining portion of the dataset, specified by the calibration ratio, is reserved for computing conformal scores, which are essential for forming prediction sets during inference.

For example, if the calibration ratio is set to 0.1, then 90% of the data is allocated for training while the remaining 10% is used for calibration. Clearly, the calibration ratio influences the interplay between training and calibration, balancing predictive performance against the reliability of uncertainty quantification. Larger calibration ratios may enhance the reliability of conformal scores, but this comes at the expense of reducing the amount of data available for training. Conversely, smaller calibration ratios favor training by allocating more data, but may compromise the quality of the calibration, and consequently, the effectiveness of the prediction sets.

B. Calibration Phase

Let $\mathcal{X}_c = \{x_{c,i}\}_{i=1}^{n}$ denote the calibration set. The calibration phase is conducted to compute conformal scores and establish the adjusted quantile threshold q_α. This process involves the following steps:

1) For each calibration point $x_{c,i} \in \mathcal{X}_c$:
   - Compute the model's output logits and apply the softmax function to obtain class probabilities.
   - Rank the classes in descending order based on their softmax scores.
   - Identify the rank of the true label $y_{c,i}$ within the sorted class probabilities.
   - Compute the conformal score E_i, which represents the minimum cumulative probability mass required to include the true label:

$$E_i = \sum_{k=1}^{\text{rank}(y_{c,i})} \hat{\pi}(x_i)_{(y_k)} \tag{1}$$

   where $\hat{\pi}(x_i)_{(y_k)}$ denotes the softmax score of the k-th ranked in descending order class.
2) Determine the adjusted quantile threshold q_α using the set of computed conformal scores and a chosen significance level α:

$$q_\alpha = \text{Quantile}(\{E_i\}_{i=1}^{n}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}) \tag{2}$$

where n is the size of the calibration set.

C. Prediction Phase

Given a test point x_t, the prediction phase uses the previously computed adjusted quantile threshold q_α to form the CP set corresponding to the point x_t. The steps involved are as follows:

1) Compute the model's output logits for x_t and apply the softmax function to obtain class probabilities.
2) Rank the classes in descending order based on their softmax scores and compute the cumulative sum of the probabilities.
3) Construct the conformal prediction set $\mathcal{C}(x_t)$, which includes the most likely classes, i.e., the classes whose cumulative probability mass is bigger than or equal to the threshold q_α:

$$\mathcal{C}(x_t) = \{y_k \mid k \in \{1, \ldots, K\}, \sum_{k=1}^{K} \hat{\pi}(x_t)_{(y_k)} \geq q_\alpha\}, \tag{3}$$

where $\hat{\pi}(x_t)_{(y_k)}$ represents the softmax score of the k-th class ranked in descending order, and q_α is the quantile threshold.

This process is repeated for each test sample, resulting in a CP set for every instance. The length of these prediction sets serves as a calibrated measure of confidence in the predictions of the model, allowing a comprehensive evaluation of its predictive performance and uncertainty.

D. Proposed Metric: CPCF

The key contribution of this work is the introduction of the CPCF as a novel metric to evaluate CF in CL settings using CP sets.

The central idea is to leverage CP sets, which quantify the model's uncertainty, to evaluate how well the model retains knowledge from previously learned tasks after training on a new task. By examining the average length of prediction sets corresponding to previously learned tasks, the CPCF effectively captures the model's confidence in its prior knowledge. A shorter average prediction set indicates higher confidence and stronger retention, while longer sets suggest increased uncertainty and potential forgetting.

Specifically, to compute the CPCF after training the model on task j, we follow the steps outlined in Algorithm 1.

After training the model on the new task j, we use the calibration sets from all previous tasks, i.e., up to task $j-1$ to compute the corresponding conformal scores $\{E_i\}_{i=1}^{l_j}$, where $l_j = \sum_{k=1}^{j-1} n_k$ and n_k denote the size of the calibration set for task k. Then, we determine the corresponding quantile threshold q_α^j to construct the prediction set $\mathcal{C}(x_t)$ for each test sample from the test sets of the previous tasks, i.e., up to task $j-1$. The length of each resulting prediction set is computed, and the CPCF is obtained as the average prediction set length across all test samples up to the previous task. This process quantifies the model's confidence in its prior knowledge and serves as a quantitative measure of CF of previous tasks after training the model on a new task j.

Algorithm 1 Computing CPCF after Training on a new Task

Input: Trained model up to task $j-1$, training data for task j, calibration set for tasks 1 to $j-1$, i.e., $\mathcal{X}_c^{1:j-1}$, and test data for tasks 1 to $j-1$, i.e., $\mathcal{X}_t^{1:j-1}$.
Output: CPCF_j assessing the CF of previous tasks, i.e., tasks $1, 2, \ldots, j-1$.
Step 1: Train the Model on Task j
Step 2: Compute Conformal Scores of Previous Tasks Calibration Samples
for each calibration sample $x_{c,i} \in \mathcal{X}_c^{1:j-1}$ **do**
    Compute the conformal score E_i.
Step 3: Compute the Quantile Threshold q_α^j
Determine the quantile threshold q_α^j:

$$q_\alpha^j = \text{Quantile}(\{E_i\}_{i=1}^{l_j}, \frac{\lceil (l_j+1)(1-\alpha) \rceil}{l_j}),$$

where $l_j = |\mathcal{X}_c^{1:j-1}|$
Step 4: Form Prediction Sets of Previous Tasks Test Samples
for each test sample $x_t \in \mathcal{X}_t^{1:j-1}$ **do**
    Form the prediction set $\mathcal{C}(x_t)$
Step 5: Compute CPCF assessing CF of Tasks $1, \ldots, j-1$

$$\text{CPCF}_j = \frac{1}{|\mathcal{X}_t^{1:j-1}|} \sum_{x_t \in \mathcal{X}_t^{1:j-1}} |\mathcal{C}(x_t)|,$$

Return: CPCF_j.

The CPCF metric provides a quantifiable approach to assess CF by observing changes in the lengths of CP sets. Shorter prediction set lengths indicate higher confidence and better retention of prior knowledge, whereas longer lengths suggest increased uncertainty, potentially resulting from forgetting previously learned tasks.

III. Experimental Results

A. Datasets Used

We evaluate our approach on four benchmark datasets: MNIST, CIFAR-10, KMNIST, and FashionMNIST. These datasets encompass a diverse set of visual classification tasks, ranging from simple digit recognition (MNIST, KMNIST) to more complex object recognition (CIFAR-10, FashionMNIST), and include both grayscale and color images.

B. Model Architecture

We employ a three-layer Multi-Layer Perceptron (MLP) for all experiments. It processes flattened image inputs, with a dimension of 784 for MNIST, KMNIST, and FashionMNIST, and 1024 for CIFAR-10. The network consists of two hidden layers with 256 and 128 ReLU-activated neurons, respectively. These are followed by a 10-neuron output layer (one neuron

per class), with a softmax activation function applied to generate class probabilities.

C. Training Curriculum

During training, the model learns a mapping $f_\theta : \mathcal{X} \to \mathcal{Y}$, where θ denotes the model parameters. Taking into account the CL setting based on incremental learning [18], the training process occurs in two main phases.

1) **Base Training:** The model is initially trained for 8 epochs in the base task, which involves learning to classify the first set of classes $\{0, 1, 2, 3, 4\}$. This phase establishes a foundational representation of the data.
2) **Incremental Training:** The model is subsequently updated incrementally to learn new tasks, with 3 epochs allocated to learning each of the remaining 5 classes. During this phase, the model aims to incorporate new knowledge while retaining previously learned information, thereby mitigating CF.

The extended base training period, relative to the shorter incremental task training phases, enables the model to build a robust foundational representation for the initial set of tasks. In contrast, the reduced number of epochs during incremental learning highlights the challenges of CL, where the model must efficiently adapt to new tasks while simultaneously preserving the knowledge of previously learned ones.

The learning rate was set at 2×10^{-5}, a conservative choice aimed at balancing rapid adaptation with long-term retention. This training setup aligns with the principle *"learn fast, forget fast; learn slow, forget slow"*, as illustrated in Figure 2, where a slower learning rate helps mitigate CF. By adopting a gradual learning pace, the model retains prior knowledge more effectively while steadily integrating new information.

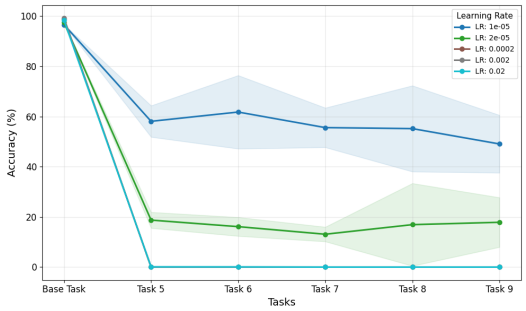

Fig. 2: Impact of learning rate on CF in MNIST. The y-axis represents the accuracy on previously learned tasks.

D. Evaluation Metrics

To assess the model's ability to retain knowledge from previous training sessions, we adopt evaluation metrics inspired by [18], which provide a robust baseline for evaluating CF. We begin by defining the two key metrics used throughout this study.

- **Accuracy on previously learned tasks** (a_{prev}): The average test accuracy over all tasks encountered prior to the current one. Specifically, $a_{\text{prev},j}$ denotes the average test accuracy on tasks 1 to $j-1$, measured after training on task j.
- **Accuracy on newly learned tasks** (a_{new}): The test accuracy on the most recently learned task j.

To show the effectiveness of the proposed CPCF in detecting and quantifying CF, we compare it with a_{prev}, and empirically show a strong correlation between CPCF and a_{prev}.

E. Numerical Results

By analyzing the relationship between CPCF and a_{prev}, we demonstrate the effectiveness of the metric in capturing the dynamics of forgetting previously learned knowledge. In addition, we investigate the robustness of CPCF to hyperparameter variations, showcasing its versatility across diverse learning scenarios. To further evaluate the reliability of CPCF in detecting CF, we compare results across two training paradigms: a standard MLP trained with cross-entropy loss, and the same MLP model trained with the EWC loss. By comparing CPCF's correlation with a_{prev} in both settings, we assess whether CPCF remains consistent and sensitive in the presence of a forgetting mitigation strategy.

We evaluated two variants of the EWC to explore how different regularization strategies influence CF and model confidence. The core idea behind EWC is to mitigate CF by selectively constraining the parameters that are most important to previously learned tasks. After training on a task A, the model estimates the importance of each parameter by computing the diagonal of the Fisher Information Matrix F, and stores the corresponding optimal parameters θ^*_A. This leads to the following regularized loss function when training on a subsequent task B:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta^*_{A,i})^2. \tag{4}$$

This variant, which we refer to as *single-penalization EWC*, applies a quadratic penalty that anchors the model parameters toward the solution found for task A. The hyperparameter λ controls the strength of this regularization, balancing stability (retention of prior knowledge) and plasticity (adaptation to new tasks).

When multiple tasks are introduced incrementally, the standard single-penalization EWC tends to primarily emphasize the most recently learned task. Thus, a *multi-penalization EWC* variant that explicitly regularizes the model with respect to all previously learned tasks could be more effective for mitigating CF. This is achieved by maintaining a separate Fisher information matrix $F^{(j)}$ and the corresponding optimal parameters θ^*_j for each previous task j. The resulting loss function for training on task t incorporates multiple regularization terms, one for each previous task:

$$\mathcal{L}(\theta) = \mathcal{L}_t(\theta) + \sum_{j=1}^{t-1} \frac{\lambda_j}{2} \sum_i F_i^{(j)}(\theta_i - \theta^*_{j,i})^2. \tag{5}$$

Each term in the summation serves as an individual constraint aimed at preserving knowledge from a specific past

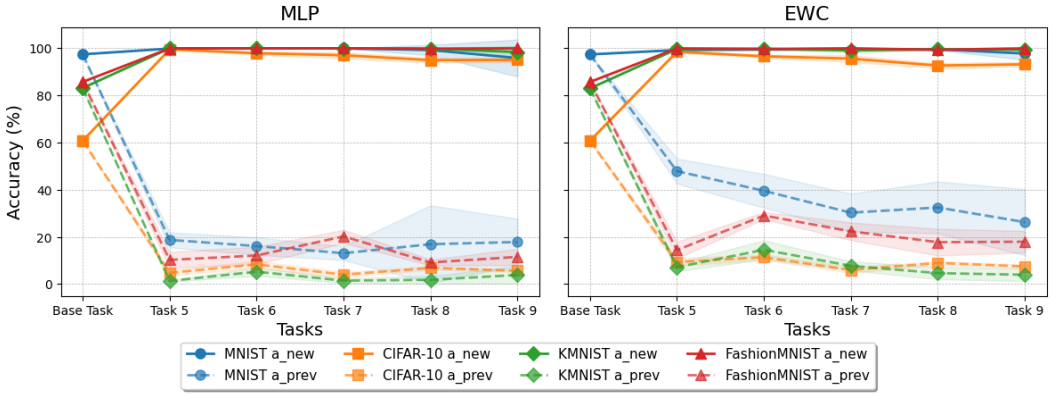

Fig. 3: Illustration of Catastrophic Forgetting: Comparison of accuracy on previously learned tasks (a_{prev}) and newly learned tasks (a_{new}) across incremental tasks.

task. This formulation offers stronger protection against forgetting by maintaining proximity to multiple past optima, though at the cost of increased memory requirements. To balance the influence of earlier tasks, we set $\lambda_j = \frac{\lambda}{2^{t-j-1}}$, assigning progressively greater weight to more recent tasks. This decay reflects the intuition that each task implicitly carries knowledge from earlier ones, thereby reducing the need to equally penalize distant past tasks.

To compare the two variants, we adopt the evaluation framework of [18], which introduces the following three metrics:

- Ω_{new}: the model's average ability to learn and retain new tasks immediately after training,
- Ω_{all}: the average accuracy across all learned tasks at each step, normalized by the base accuracy (a_{ideal}),
- Ω_{base}: the retention of the base task after each new task is learned, normalized by a_{ideal}.

We also introduce $\Omega_{\textbf{prev}}$, which assesses the model's ability to retain previously learned tasks after learning a new one. We compute it from the accuracy of previous tasks in each step, using a_{prev}. Formally, it is defined as:

$$\Omega_{\text{prev}} = \frac{1}{T-1} \sum_{j=2}^{T} \frac{a_{\text{prev},j}}{a_{\text{ideal}}}. \tag{6}$$

TABLE I: Comparison of two EWC variants on CIFAR-10 using Ω metrics.

Method	Ω_{base}	Ω_{new}	Ω_{all}	Ω_{prev}
Single-penalization EWC	0.082	0.953	0.101	0.342
Multi-penalization EWC	0.082	0.953	0.101	0.342

As shown in Table I, both EWC variants achieve identical scores on CIFAR-10 in all evaluation metrics: Ω_{new}, Ω_{all}, Ω_{base}, and Ω_{prev}. Although the single-penalization approach only penalizes deviations from the most recently learned task, it still indirectly preserves earlier knowledge, as each updated parameter configuration reflects the cumulative training history. Therefore, to reduce computational overhead and memory requirements, we adopted the single-penalization variant, referred to simply as EWC throughout the remainder of this study.

Figure 3 highlights the phenomenon of CF, where the models studied maintain high accuracy in newly introduced tasks while exhibiting a sharp decline in accuracy performance in previously learned tasks. This drop in a_{prev} reflects the model's difficulty in retaining knowledge from earlier tasks, ultimately leading to a failure to generalize across all learned tasks. The right panel shows that EWC partially alleviates this degradation by preserving higher a_{prev} scores compared to standard MLP, particularly during the early stages of CL.

Tables II and III reveal a strong distance correlation between CPCF and a_{prev} in various calibration ratios and significance levels α. Specifically, this correlation becomes consistently stronger when EWC is applied. This observation suggests that CPCF becomes more sensitive to forgetting in the presence of CF mitigation techniques, reinforcing its value as a reliable proxy for measuring forgetting in CL scenarios.

Table II shows the distance correlation between CPCF and a_{prev} in varying calibration ratios, with α fixed at 0.1. Although some fluctuations are observed, there is no consistent trend suggesting that increasing the calibration ratio systematically improves or worsens the correlation. In some cases, a higher calibration ratio leads to marginal gains; in others, the correlation remains stable or slightly decreases. This behavior suggests that CPCF is relatively robust to the choice of calibration ratio, i.e., its alignment with forgetting, as measured by a_{prev}, does not critically depend on precise calibration tuning. In all settings, models trained with EWC consistently show stronger correlations than those trained with standard methods, further validating the sensitivity of CPCF under forgetting-aware regularization.

Table III further explores the effect of varying the significance level α. As expected, increasing α leads to larger prediction sets and stronger correlations with a_{prev}, as CPCF becomes more sensitive to subtle changes in model uncertainty. This behavior aligns with the theoretical role of α in CP: it directly determines the quantile threshold used to construct prediction sets, thus influencing their size and informativeness.

TABLE II: Distance correlation between CPCF and a_{prev} for fixed alpha ($\alpha = 0.1$).

Calib.	Model	MNIST	CIFAR-10	FashionMNIST	KMNIST
0.05	MLP	0.5165	0.3009	0.2319	0.3214
	EWC	0.7475	0.3848	0.2993	0.5970
0.10	MLP	0.5585	0.3211	0.2118	0.3121
	EWC	0.6754	0.3883	0.3046	0.5996
0.15	MLP	0.5096	0.3363	0.1863	0.2709
	EWC	0.7081	0.4414	0.3149	0.5693
0.20	MLP	0.4996	0.3319	0.1956	0.2655
	EWC	0.7127	0.4834	0.2903	0.5298

TABLE III: Distance correlation between CPCF and a_{prev} for fixed calibration ratio (calibration ratio $= 0.1$).

Alpha	Model	MNIST	CIFAR-10	FashionMNIST	KMNIST
0.05	MLP	0.5347	0.2871	0.2611	0.2321
	EWC	0.6250	0.2574	0.2740	0.5812
0.10	MLP	0.5585	0.3211	0.2118	0.3121
	EWC	0.6754	0.3883	0.3046	0.5996
0.15	MLP	0.5732	0.3202	0.1834	0.3116
	EWC	0.7480	0.4270	0.3073	0.6106
0.20	MLP	0.5919	0.3224	0.1777	0.3107
	EWC	0.8012	0.4540	0.3157	0.6201

A higher α corresponds to a lower threshold, resulting in larger prediction sets that are more likely to contain the true label but may also be less precise. These wider sets more effectively capture the model's uncertainty, especially in regions of the input space where the model is less confident. Consequently, CPCF becomes increasingly sensitive to subtle degradations in model performance, particularly those introduced by forgetting previously learned tasks. This sensitivity explains the observed increase in correlation with a_{prev} at higher α values. In general, this behavior highlights an important strength of CPCF: its adaptability to different levels of uncertainty tolerance. By tuning α, practitioners can control the granularity with which forgetting is detected, making CPCF a versatile tool for evaluating CL performance.

IV. Discussion and Concluding Remarks

The proposed method presents a novel application of CP to quantify CF in CL, introducing CPCF as an effective and interpretable online metric.

By capturing the relationship between model uncertainty and knowledge retention, CPCF offers valuable insights into the dynamics of forgetting. Furthermore, the method adheres to the principles of *trustworthy AI* by promoting transparency, reliability, and confidence in decision-making processes.

While the current study focuses on classification, the proposed framework can naturally be extended to regression tasks by transforming CP prediction sets into CP interval estimates. This generalization broadens the scope of the CPCF, making it applicable to a wider range of learning problems while preserving its interpretability and effectiveness.

Future work may focus on optimizing CPCF for greater efficiency and exploring its utility in critical domains such as medical diagnostics, autonomous systems, and cybersecurity, where managing uncertainty and mitigating CF are vital. Additionally, extending the approach to handle diverse and complex data distributions, including multimodal, imbalanced, or non-stationary data streams, would further enhance its real-world applicability.

Acknowledgments

The work has been supported by the EC through the Horizon Europe/JU SNS project, ROBUST-6G (Grant Agreement no. 101139068). N. Jamoussi is also partially funded by the IMT "Futur, Ruptures & Impacts" program.

References

[1] G. Serra, B. Werner, and F. Buettner, "How to leverage predictive uncertainty estimates for reducing catastrophic forgetting in online continual learning," 2024. [Online]. Available: https://arxiv.org/abs/2407.07668

[2] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, 2024.

[3] C. V. Nguyen, A. Achille, M. Lam, T. Hassner, V. Mahadevan, and S. Soatto, "Toward understanding catastrophic forgetting in continual learning," 2019. [Online]. Available: https://arxiv.org/abs/1908.01091

[4] A. Shahid *et al.*, "Large-scale AI in telecom: Charting the roadmap for innovation, scalability, and enhanced digital experiences," *arXiv:2503.04184*, 2025.

[5] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," ser. Psy. of Le. and Mot. Academic Press, 1989, vol. 24, pp. 109–165.

[6] G. D. Y, N. G. Nair, P. Satpathy, and J. Christopher, "Covariate shift: A review and analysis on classifiers," in *2019 Global Conference for Advancement in Technology (GCAT)*, 2019.

[7] Z. C. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," 2018. [Online]. Available: https://arxiv.org/abs/1802.03916

[8] V. Vovk, "Testing for concept shift online," 2020. [Online]. Available: https://arxiv.org/abs/2012.14246

[9] J. Kirkpatrick *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the Nat. Ac. of Sc.*, vol. 114, no. 13, p. 3521–3526, Mar. 2017.

[10] H. Ritter *et al.*, "Online structured laplace approximations for overcoming catastrophic forgetting," 2018. [Online]. Available: https://arxiv.org/abs/1805.07810

[11] J. S. Vitter, "Random sampling with a reservoir," *ACM Trans. Math. Softw.*, vol. 11, no. 1, p. 37–57, Mar. 1985. [Online]. Available: https://doi.org/10.1145/3147.3165

[12] L. Wang *et al.*, "A comprehensive survey of continual learning: Theory, method and application," 2024. [Online]. Available: https://arxiv.org/abs/2302.00487

[13] B.-S. Einbinder *et al.*, "Training uncertainty-aware classifiers with conformalized deep learning," 2022. [Online]. Available: https://arxiv.org/abs/2205.05878

[14] I. Pitsiorlas, G. Arvanitakis, and M. Kountouris, "Trustworthy intrusion detection: Confidence estimation using latent space," in *2024 22nd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2024, pp. 92–98.

[15] I. Pitsiorlas, A. Tsantalidou, G. Arvanitakis, M. Kountouris, and C. Kontoes, "A latent space metric for enhancing prediction confidence in earth observation data," in *2024 IEEE Intern. Geoscience and Remote Sensing Symp. (IGARSS 2024)*, 2024, pp. 987–991.

[16] V. K. Kurmi *et al.*, "Do not forget to attend to uncertainty while mitigating catastrophic forgetting," 2021. [Online]. Available: https://arxiv.org/abs/2102.01906

[17] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, vol. 16, no. 4, p. 494–591, Mar. 2023.

[18] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *Proceedings of the 32nd AAAI Conf. on AI and 30th Innov. App. of AI Conf. and 8th AAAI Symp. on Educ. Advances in AI*. AAAI Press, 2018.