Reliable TV Audience Forecasting with Conformal Prediction

Javier Carreno[1], Khuong An Nguyen[1], Zhiyuan Luo[1], and Andrew Fish[2]

[1] Royal Holloway University of London, Surrey TW20 0EX, United Kingdom
Javier.Carreno.2023@live.rhul.ac.uk
{Khuong.Nguyen, Zhiyuan.Luo}@rhul.ac.uk
[2] University of Liverpool, Liverpool L69 3BX, United Kingdom
Andrew.Fish@liverpool.ac.uk

Abstract. Accurately identifying a household demographic profile based on its television viewing pattern is important for content personalisation, targeted advertising, and programme design. By understanding who is watching what and when, broadcasters can tailor content to match viewers' interests. Although machine learning can predict household attributes, uncertainty is often high due to overlapping viewing patterns across demographic groups, shared device usage, and limited samples. Thus, this paper applies the Conformal Prediction framework to provide an uncertainty measure for machine prediction. We also introduce a new nonconformity score to improve prediction efficiency. Experiments on a large-scale, imbalanced TV dataset show that our method achieves an average prediction set size of 1.18 and an 82.8% singleton rate at 95% confidence level, outperforming conventional nonconformity measures in terms of both reliability and efficiency.

Keywords: Conformal Prediction · Nonconformity Measure · TV household classification

1 Introduction

In recent years, AI has become a powerful tool for extracting actionable insights from user behaviour across digital and traditional media platforms [12]. In the context of linear television, accurate audience segmentation and demographic profiling are essential for content personalisation, targeted advertising, and strategic programming decisions. In our previous work [6], we developed a machine learning framework capable of predicting household demographics based on viewing behaviour. However, the framework lacked a quantifiable measure of confidence in its predictions—a critical limitation for real-world deployment [3].

The absence of confidence estimates is particularly problematic in applications where interpretability and decision reliability are essential. Broadcasters and advertisers must often balance precision with uncertainty, especially when dealing with heterogeneous audiences and imbalanced data distributions. In such settings, producing well-calibrated, interpretable prediction sets—rather than single-label outputs—improves audience targeting and scheduling decisions.

Conformal Prediction (CP) addresses the challenge of uncertainty in classification by enabling models to output prediction sets—collections of likely labels—instead of committing to a single prediction. By offering formal guarantees on coverage (e.g., ensuring the true label is included in the set with high probability), CP allows practitioners to control the risk of error in a transparent and interpretable way [13,20]. These properties make CP a powerful tool for building more reliable machine learning systems.

At the core of CP lies the Nonconformity Measure (NCM), which quantifies how 'atypical' a new instance is relative to previously seen data. While a range of NCMs—based on predicted probabilities, rankings, or uncertainty—are commonly used, they are typically applied uniformly across instances, disregarding instance-specific difficulty or context [4]. This uniformity can lead to suboptimal performance. Some NCMs may yield efficient prediction sets for well-classified instances but fail on ambiguous or rare cases. To address this, we introduce a novel calibration-aware NCM that integrates instance-level accuracy with global model calibration, enabling more robust and context-sensitive prediction sets.

The key contributions of this paper are as follows:

- We propose a calibration-aware alternative to standard NCMs, combining local prediction accuracy with global reliability, and evaluate its performance under varying conditions.
- We empirically evaluate our method on real-world data across a set of metrics designed to assess both validity and efficiency. Validity refers to the probability that the prediction set contains the true value, while efficiency relates to the size of the prediction set.

The remainder of the paper is organised as follows. Section 2 reviews relevant work on demographic inference and conformal prediction. Section 3 formalises the household classification task and motivates the need for uncertainty-aware predictions. Section 4 introduces the conformal prediction framework and reviews key nonconformity measures. Section 5 presents our proposed Hybrid Calibration Score (HCS). Section 6 details the experimental design, dataset, evaluation metrics, and results. Finally, Section 7 concludes and outlines future directions.

2 Related Work

Demographic inference from user behaviour is a long-standing challenge in audience research, with applications in marketing, recommendation systems, and audience measurement. Prior studies have drawn on diverse data sources—such as web browsing logs, mobile app usage, and television viewing patterns—to predict attributes like age, gender, household composition, and interests [15,19]. Approaches have evolved from rule-based heuristics to supervised learning methods, including decision trees, random forests, and deep neural networks.

Television viewership, in particular, has been a valuable source for household profiling. Our prior work advanced this area by applying supervised learning to

large-scale, first-party linear TV data to predict household demographics with high accuracy [6]. However, these models lacked mechanisms to quantify uncertainty, limiting their interpretability and reliability in deployment.

CP addresses this limitation by producing statistically valid prediction sets under minimal distributional assumptions [20]. While CP has gained traction in areas such as medical diagnostics [18], recommender systems [17], and natural language processing [9, 10], its use in behavioural media analytics remains limited.

Most CP methods rely on conventional NCMs applied uniformly across instances, ignoring instance-specific difficulty or uncertainty [3, 11]. Recent work aims to improve contextual sensitivity through instance-aware NCMs. Seedat et al. [16] integrate CP with self-supervised learning (SSL), using errors from an auxiliary task as uncertainty signals. Bellotti [5] proposes a differentiable CP framework that optimises task-specific loss functions to produce tighter, context-aware prediction intervals. Similarly, Amoukou and Brunel [2] use Quantile Regression Forests (QRFs) to generate localised intervals by weighting calibration residuals based on feature-space similarity.

These advances reflect a broader shift toward adaptive, learnable NCMs that go beyond global heuristics. However, many require task-specific adaptations, auxiliary training signals, or model-dependent architectures. In contrast, we explore model-agnostic alternatives that incorporate uncertainty and calibration directly into the NCM, without modifying the underlying learning pipeline.

Unlike prior approaches such as SSL-based CP [16] or differentiable CP [5], which rely on auxiliary training objectives or architectural changes, the method proposed in this paper is fully model-agnostic. It operates directly on model output probabilities and can be applied without retraining or internal access to the model—making it particularly suitable for deployment in production settings or black-box environments.

3 Demographic Prediction Problem

We address the task of predicting household demographics from patterns of TV viewing behaviour—a challenge especially relevant in linear TV, where demographic data is often unavailable but vital for audience segmentation, content scheduling, and advertising.

Let $\mathcal{X}$ denote the space of behavioural features from first-party TV consumption data, and $\mathcal{Y}$ a finite set of six demographic classes representing distinct stages in the household life cycle (e.g., '*Only middle-aged adults*', '*Seniors*'; see Table 1). Each instance consists of a feature vector $x_i \in \mathcal{X}$ and a label $y_i \in \mathcal{Y}$. The goal is to learn a classifier $f : \mathcal{X} \to \mathcal{Y}$ that generalises to new households.

Prior work shows that supervised models can classify these categories with high accuracy using features like time-of-day activity and content preferences [6]. However, they yield single-label predictions without uncertainty—limiting their reliability when viewing patterns are noisy or ambiguous.

To overcome this, we apply *Conformal Prediction (CP)*. CP augments any base classifier to produce calibrated prediction sets $\Gamma(x) \subseteq \mathcal{Y}$ that contain the true label with a user-specified confidence level (e.g., 95%). These prediction sets deliver calibrated uncertainty estimates, especially useful for ambiguous or underrepresented cases.

4 Conformal Prediction Background

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output (label) space. Given a dataset of n i.i.d. examples,

$$Z_{1:n} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X}^n \times \mathcal{Y}^n,$$

CP provides a framework for constructing prediction sets $\Gamma(x_{n+1}) \subseteq \mathcal{Y}$ for a new input $x_{n+1} \in \mathcal{X}$, such that the true label $y_{n+1} \in \mathcal{Y}$ is included with a pre-specified probability $1 - \alpha$:

$$\mathbb{P}(y_{n+1} \in \Gamma(x_{n+1})) \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$ is the user-defined significance level. This guarantee holds under the assumption of exchangeability, which only requires that the joint probability distribution of n examples does not change when the order of those examples is changed.

4.1 Nonconformity Measure (NCM)

A core component of CP is the *Nonconformity Measure* (NCM),

$$A : \mathcal{Z}^n \times \mathcal{Z} \to \mathbb{R},$$

which assigns a scalar score. Given a nonconformity measure (A_{n+1}) and a bag $\wr z_1, ..., z_n \int$ of n training examples, we can compute the nonconformity score $a_i(x_i, y_i) = A_{n+1}(\wr z_1, ..., z_{i-1}, z_{i+1}, ..., z_{n+1} \int, z_i)$ indicating how atypical each pair $z_i = (x_i, y_i)$ is relative to the training data of n examples in the bag. In ICP, this score is computed using the trained model and applied across calibration and test instances.

The choice of NCM is critical [14], as it directly influences the size, informativeness, and adaptiveness of prediction sets. Broadly, NCMs can be categorised as either *model-dependent* or *model-agnostic* [1]:

- *Model-dependent NCMs* rely on internal model outputs such as predicted probabilities, confidence scores, or learned embeddings. Examples include probability-based scores (e.g., log loss, hinge loss), ranking-based metrics (e.g., margin, gap), and uncertainty measures (e.g., entropy).
- *Model-agnostic NCMs* do not depend on the internal structure of the model. Instead, they use external features such as distances in input or embedding space (e.g., KNN distance, embedding distance), offering greater flexibility but often at the cost of tighter model alignment.

For clarity, we group commonly used NCMs into four operational categories: *probability-based scores*, *ranking/confidence scores*, *uncertainty metrics*, and *distance-based scores*.

a) Probability-Based Scores These scores use predicted probabilities directly:

- **Hinge Loss:** $a(x, y) = 1 - p(y|x)$
- **Log Loss:** $a(x, y) = -\log p(y|x)$
- **Brier Score:**
$$a(x, y) = \sum_{y' \in \mathcal{Y}} \left(p(y'|x) - \mathbf{1}_{\{y'=y\}}\right)^2$$

Log Loss and *Brier Score* are both **proper scoring rules**, rewarding accurate and well-calibrated probability estimates.

b) Ranking/Confidence Scores These reflect a model's confidence or ranking of labels:

- **Margin:** $a(x, y) = p(y_1|x) - p(y_2|x)$, where y_1 and y_2 are the top two predicted labels and $y_1, y_2 \in \mathcal{Y}$.
- **Gap:** $a(x, y) = p(y_1|x) - p(y|x)$, where y_1 is the top prediction.

c) Uncertainty Metrics These assess overall uncertainty in the prediction:

- **Entropy:** $a(x, y) = -\sum_{y' \in \mathcal{Y}} p(y'|x) \log p(y'|x)$

d) Distance-Based Scores These quantify instance deviation in feature space:

- **Embedding Distance:** $a(x, y) = \|\phi(x, y) - \mu\|$, where $\phi(x, y)$ is a learned embedding and μ is the mean embedding over the calibration data.
- **KNN Distance:** $a(x, y) = \frac{1}{K} \sum_{i=1, y_i=y}^{K} \|x - x_i\|$

5 Hybrid Calibration Score (HCS)

Conventional NCMs typically focus on local prediction properties, such as per-instance correctness or margin, but often neglect global model calibration. To tackle this, we propose a composite score that integrates both instance-level accuracy and global calibration quality.

The Hybrid Calibration Score (HCS) combines the *Brier Score* and *Log Loss*—both proper scoring rules applied to individual instances—with the *Expected Calibration Error (ECE)*, a global calibration metric computed once over the calibration set, as introduced by Guo et al. [7]. The weights $\alpha, \beta, \gamma \in \mathbb{R}^+$ determine the contribution of each component:

$$a(x, y) = \alpha \cdot \text{Brier}(x, y) + \beta \cdot \text{LogLoss}(x, y) + \gamma \cdot \text{ECE}, \tag{1}$$

We optimise the HCS weights using Bayesian optimisation with Gaussian Processes, implemented via `gp_minimize` from the `scikit-optimize` library. The objective penalises both large prediction sets and coverage violations relative to the target confidence level. To ensure interpretability, the weights are constrained to form a convex combination and renormalised at each step.

The search space for the individual weights is defined over the simplex:

$$\alpha, \beta, \gamma \in [0.0, 1.0] \quad \text{subject to } \alpha + \beta + \gamma = 1$$

We run up to 50 evaluations, starting from 10 random initialisations, which provides a practical trade-off that enables reliable convergence.

To check robustness, we repeat the optimisation across five different calibration/test splits. The resulting weights remain consistent across folds, and performance in terms of coverage and average prediction set is stable—suggesting the method is not overfitting to any particular split. We also apply early stopping if improvements plateau and include a regularisation term to prevent over-reliance on any single component (e.g., Brier alone). These safeguards reduce the risk of overfitting and improve robustness across data splits.

Expected Calibration Error (ECE) The *Expected Calibration Error (ECE)* quantifies the discrepancy between predicted confidence and observed empirical accuracy across varying confidence levels. Specifically, predicted probabilities are partitioned into n equally spaced bins. For each bin B_i, we compute:

- $\text{acc}(B_i)$: the empirical accuracy in bin B_i,
- $\text{conf}(B_i)$: the average predicted confidence in bin B_i.

The ECE is defined as:

$$\text{ECE} = \sum_{i=1}^{n} \frac{|B_i|}{N} \cdot |\text{acc}(B_i) - \text{conf}(B_i)| \tag{2}$$

where $|B_i|$ is the number of samples in bin i, and N is the total number of samples. The absolute difference ensures that both overconfident and underconfident predictions contribute positively to the error.

An ideally calibrated model satisfies $\text{acc}(B_i) = \text{conf}(B_i)$ for all bins i, resulting in $\text{ECE} = 0$. Higher ECE values indicate poorer calibration.

Within the proposed **HCS**, *ECE* acts as a shared global regularisation term that complements the instance-wise components. While the *Brier Score* and *Log Loss* focus on individual prediction quality, *ECE* enforces alignment between predicted confidence and actual correctness at the population level, promoting probabilistically trustworthy outputs.

6 Empirical Results

This study builds on prior work applying machine learning to classify households by television viewing behaviour [6]. Demographic labels were heuristically derived from business rules and external market insights. Though approximate, they capture key behavioural patterns across household types.

The dataset comprises 19,386 instances and exhibits a pronounced class imbalance (Table 1). Categories such as '*Only middle-aged adults*' and '*Seniors*' are overrepresented, while others—such as '*Couples with young kids*'—are rare. This skew poses challenges for both classification and CP, which depends on well-calibrated confidence scores. The dataset along with detailed documentation are publicly available at https://github.com/carrenyo/TV-Viewer-Demographics-Machine-Learning.

Table 1: Household Classification Distribution.

Household Classification	**Num of Devices**	**Percentage**
Couple with young kids (0-8 years)	143	0.74%
Couple with teenagers (9-17 years)	1,514	7.81%
Couple with adult children (18+ years)	3,567	18.39%
Only young adults (18-35 years)	3,550	18.30%
Only middle-aged adults (36+ years)	6,985	36.02%
Seniors (elderly/retired adults)	3,627	18.74%

To mitigate this imbalance, we applied stratified sampling to preserve class distributions when splitting the data, aligning with findings that the size of the calibration set significantly impacts the validity of conformal prediction [3].

Among the models evaluated in our previous work—*Random Forest*, *K-Nearest Neighbours*, and *Gradient Boosting*—we selected **Random Forest** as the primary classifier due to its balanced performance across demographic groups, as well as its stability and reliability during cross-validation.

6.1 Evaluation Metrics

To assess the performance of the conformal predictors, we use a set of widely adopted metrics [8] that jointly evaluate the *validity* and *efficiency* of the prediction sets:

- **Coverage:** The proportion of test instances for which the true class is included in their prediction set. A conformal predictor is considered valid if its empirical coverage aligns with the target confidence level.
- **Average Prediction Set Size (APS):** The mean number of labels in the prediction sets. Lower values indicate more efficient and informative predictions.

– **Singleton Rate (OneC):** The percentage of test instances for which the prediction set contains exactly one label. A higher singleton rate reflects the model's ability to make confident, and unambiguous predictions.

6.2 Experimental Setup

We evaluate the conformal prediction framework on the household classification task using a *Random Forest* classifier. The dataset is split into 60% training, 30% calibration, and 10% test sets, with stratified sampling to preserve class distributions. Results are averaged over five independent runs with different random seeds (13, 25, 45, 50, 56) to ensure robustness.

Confidence Levels. Performance is evaluated across a range of significance levels from 0.01 to 0.25, corresponding to confidence levels from 99% to 75%. This range reflects practical trade-offs between reliability and informativeness. Low confidence levels are generally too permissive for risk-sensitive applications and are excluded from consideration.

Nonconformity Measures. We evaluate four nonconformity measures: Hinge Loss, Gap, Brier Score, and our proposed Hybrid Calibration Score (HCS). These cover a range of scoring strategies introduced in Section 4.1, from confidence-based to calibration-aware. The selection reflects both diversity and empirical relevance; other measures, such as Entropy and Distance-Based Scores, were excluded due to poor decisiveness or limited compatibility with our model setup.

Implementation. All experiments are implemented in Python. Model training and prediction use `scikit-learn`. To enable flexible experimentation, we developed a lightweight wrapper that mimics the `mapie` API, extending support to custom nonconformity measures not included in the original library.

6.3 Empirical Performance

This section presents empirical results for four NCMs evaluated. Performance is assessed using three metrics: Coverage, Average Prediction Set Size (APS), and Singleton Rate (OneC). Table 2 summarises results at a significance level of $\alpha = 0.01$, corresponding to a 99% confidence level. All metrics are reported as *mean* and *standard deviation* across five random seeds to account for variability.

As shown in Table 2, all NCMs achieve near-nominal coverage at the 99% confidence level. *Hinge Loss* yields the highest mean coverage (99.32%) with minimal variance but produces relatively large prediction sets (APS = 1.64) and a moderate singleton rate. *Gap* and *Brier Score* offer comparable coverage (99.24% and 99.21%), though with larger prediction sets and slightly lower decisiveness. In contrast, *HCS* achieves the smallest prediction sets (APS = 1.55) and the highest singleton rate (55.19%), indicating greater efficiency and confidence, despite a slight drop in coverage (98.99%). Overall, *HCS* offers the best trade-off between reliability and informativeness.

Table 2: Performance Comparison of NCMs at $\alpha = 0.01$.

	Coverage (%)		**APS**		**OneC (%)**	
NCM	Mean	Std	Mean	Std	Mean	Std
Hinge Loss	99.32	0.08	1.64	0.04	49.23	2.95
Gap	99.24	0.28	2.74	0.06	54.30	1.46
Brier Score	99.21	0.29	2.50	0.08	54.78	1.67
HCS	98.99	0.17	**1.55**	0.04	**55.19**	2.71

While these results are promising at a fixed confidence level, a key question remains: do these trends hold across a broader range of confidence levels? To explore this, we extend the analysis from 99% down to 75% confidence and first assess whether each NCM satisfies the coverage guarantees required by CP.

For each confidence level, we compute the empirical coverage and assess whether it meets the expected reliability. A method is classified as:

- **Compliant** if the empirical coverage exceeds the lower bound of a one-sided 99.99% confidence interval by at least 0.004;
- **Marginal** if the empirical coverage lies within 0.004 above the lower bound;
- **Non-compliant** if it falls below the lower bound.

The lower bound is computed using the normal approximation as:

$$\text{Lower CI bound} = (1-\alpha) - 3.719 \cdot \sqrt{\frac{(1-\alpha)\alpha}{n}} \tag{3}$$

where α is the significance level and n is the number of test instances.

Figure 1 presents the empirical coverage achieved by each NCM across a range of confidence levels. All methods closely follow the ideal calibration curve, demonstrating strong overall calibration. *HCS* tends to produce slightly conservative coverage at lower confidence levels. As confidence increases, its coverage aligns more closely with the nominal rate and occasionally dips just below it, yet remains within strict compliance bounds. In contrast, *Hinge Loss*, *Gap*, and *Brier Score* consistently maintain compliant coverage across all evaluated levels.

These results confirm that all methods uphold the theoretical guarantees of conformal prediction.

Although *HCS* occasionally dips just below the nominal coverage at the highest confidence levels, this behaviour reflects a trade-off inherent in its design. By optimising for smaller prediction sets and higher singleton rates, HCS introduces a mild tolerance for under-coverage in ambiguous cases—especially those with high class overlap or low model confidence. This is a common effect of prioritising efficiency in nonconformity measures. Potential mitigations include conservative calibration adjustments (e.g., applying a small quantile shift) or incorporating temperature scaling [7] to improve probability calibration before

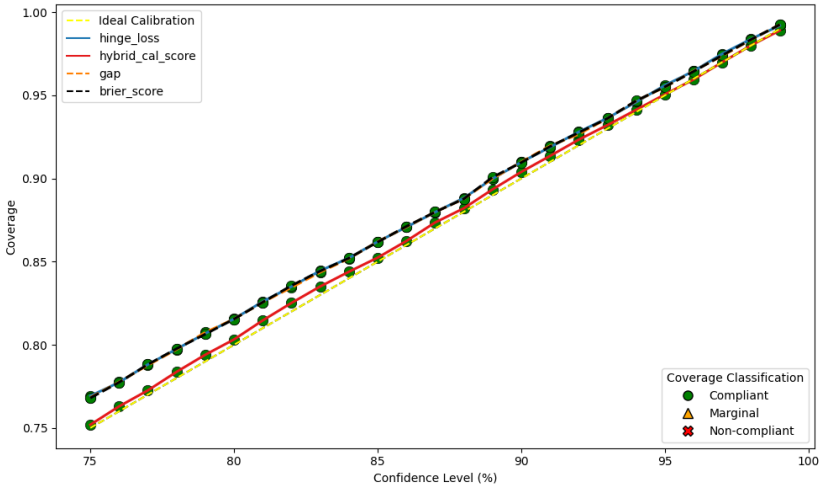

Fig. 1: Empirical Coverage vs. Nominal Confidence Level.

applying conformal prediction. These adjustments could further reduce under-coverage without compromising the decisiveness that HCS offers, and represent promising directions for future work.

Efficiency Analysis. With coverage compliance verified, we now analyse the efficiency of each NCM in terms of prediction set size and the frequency of singleton predictions.

Figure 2 shows the performance of various NCMs across confidence levels, using Average Prediction Set size (APS, top) and Singleton Frequency (OneC, bottom). *HCS* consistently yields the smallest prediction sets, indicating superior efficiency. While *Hinge Loss* performs similarly at lower confidence levels, its efficiency drops as confidence increases. Both *Gap* and *Brier Score* exhibit a sharp rise in APS beyond 95%, with *Gap* being the least efficient at 99%.

The bottom plot complements this with the *OneC* metric—the proportion of singleton prediction sets—which reflects decisiveness. All methods peak around 89% confidence, where the balance between reliability and informativeness is most favourable. *Hinge Loss* achieves the highest OneC at lower confidence levels, reflecting aggressive set reduction, but suffers the steepest decline beyond 94%, falling below all other methods at high confidence levels.

By contrast, *HCS* begins more conservatively—producing fewer singletons at low confidence—but achieves a higher OneC from 94% onward. This suggests that while *Hinge Loss* is more decisive at relaxed thresholds, *HCS* remains effective under stricter confidence requirements. *Gap* and *Brier Score* briefly outperform *HCS*, but degrade more rapidly as confidence increases.

Together, these results highlight the strength of *HCS* as a **well-balanced NCM**: it combines **high efficiency** (low APS) with **robust decisiveness**

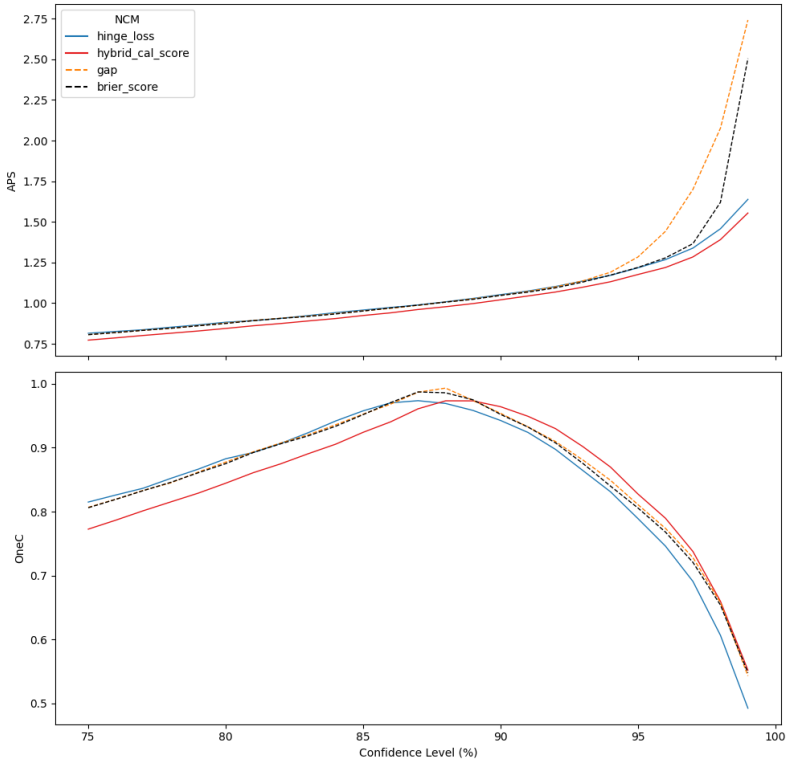

Fig. 2: APS and OneC Across Confidence Levels.

(high OneC) across a wide confidence range. Compared to the other baselines, *Hinge Loss* delivers the second-best efficiency but loses decisiveness under high-confidence constraints. *Brier Score* offers a more stable middle ground, maintaining better OneC than *Gap*, which—despite peaking highest near 89%—becomes the least efficient option as confidence increases.

To further assess the performance of the two best NCMs in different demographic groups, we compare *Hinge Loss* and *HCS* across household categories at $\alpha = 0.01$ (99% confidence). Table 3 summarises coverage, APS, and OneC.

HCS consistently yields smaller prediction sets and higher singleton rates across all categories, confirming its superior efficiency and decisiveness. The difference is most pronounced in classes with greater uncertainty—such as '*Couple with young kids*'—which remain challenging for both NCMs, with coverage falling notably below the target. In these cases, *Hinge Loss* shows a steeper decline in OneC. These findings align with the overall analysis, confirming that *HCS* offers more stable and efficient performance across diverse household categories.

Table 3: Per-Class Performance Comparison of *Hinge Loss* and *HCS* at $\alpha = 0.01$.

Category	Coverage (%)		APS		OneC (%)	
	Hinge (std)	HCS (std)	Hinge (std)	HCS (std)	Hinge (std)	HCS (std)
Couple w/ adult children	98.26 (0.46)	97.42 (0.83)	2.26 (0.06)	**2.10** (0.08)	11.43 (2.56)	**19.16** (3.97)
Couple w/ teenagers	98.54 (0.55)	98.01 (0.81)	2.25 (0.06)	**2.11** (0.05)	12.32 (2.59)	**19.74** (3.58)
Couple w/ young kids	94.29 (3.19)	94.29 (3.19)	2.69 (0.28)	**2.60** (0.19)	2.86 (3.91)	**4.29** (3.91)
Only middle-aged adults	99.77 (0.16)	99.63 (0.26)	1.34 (0.04)	**1.28** (0.04)	66.92 (4.09)	**72.42** (3.35)
Only young adults	100.00 (0.00)	100.00 (0.00)	1.63 (0.04)	**1.53** (0.04)	37.30 (4.41)	**46.93** (4.26)
Seniors	99.34 (0.31)	98.95 (0.45)	1.31 (0.05)	**1.28** (0.05)	81.16 (3.16)	**82.26** (2.91)

7 Conclusions and Future Work

This paper applies Conformal Prediction to demographic classification from television viewership data, addressing the need for reliable confidence estimates in audience segmentation. We propose the **Hybrid Calibration Score (HCS)**, a nonconformity measure combining instance-level accuracy with model calibration. Experiments on a large, imbalanced dataset show that HCS achieves strong efficiency while satisfying coverage guarantees across diverse confidence levels.

Future work includes extending *HCS* to multi-label and regression tasks, and exploring adaptive weighting in place of fixed scoring. Another direction is to evaluate *HCS* in high-stakes domains—such as healthcare, finance, or marketing—where calibrated uncertainty is critical.

Furthermore, we plan to investigate calibration refinement techniques (e.g., temperature scaling or conservative quantile adjustments) to further reduce the minor under-coverage observed at high confidence levels, while preserving the efficiency gains demonstrated by HCS.

References

1. Aleksandrova, M., Chertov, O.: Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study. In: Conformal and Probabilistic Prediction and Applications. pp. 151–170. PMLR (2021)
2. Amoukou, S.I., Brunel, N.J.: Adaptive conformal prediction by reweighting nonconformity score. arXiv preprint arXiv:2303.12695 (2023)
3. Angelopoulos, A.N., Bates, S., et al.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023)
4. Balasubramanian, V., Ho, S.S., Vovk, V.: Conformal prediction for reliable machine learning: theory, adaptations and applications. Newnes (2014)
5. Bellotti, A.: Constructing normalized nonconformity measures based on maximizing predictive efficiency. In: Conformal and Probabilistic Prediction and Applications. pp. 41–54. PMLR (2020)
6. Carreno, J., An Nguyen, K., Luo, Z., Fish, A.: Unlocking viewer insights in linear television: A machine learning approach. In: International Conference on Business Informatics Research. pp. 53–67. Springer (2024)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

8. Kato, Y., Tax, D.M., Loog, M.: A review of nonconformity measures for conformal prediction in regression. Conformal and Probabilistic Prediction with Applications pp. 369–383 (2023)
9. Lei, J., Wang, J.: Distribution-free uncertainty quantification for classification tasks with nlp applications. Transactions of the Association for Computational Linguistics **10**, 100–116 (2022)
10. Nguyen, K.A.: A performance guaranteed indoor positioning system using conformal prediction and the wifi signal strength. Journal of Information and Telecommunication **1**(1), 41–65 (2017)
11. Nguyen, K.A., Luo, Z.: Reliable indoor location prediction using conformal prediction. Annals of Mathematics and Artificial Intelligence **74**(1), 133–153 (2015)
12. Nixon, L., Ciesielski, K., Philipp, B.: Ai for audience prediction and profiling to power innovative tv content recommendation services. In: Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery. pp. 42–48 (2019)
13. Papadopoulos, H.: Inductive conformal prediction: Theory and application to neural networks. In: Tools in Artificial Intelligence. IntechOpen (2008)
14. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13. pp. 345–356. Springer (2002)
15. Razavi, R., Xue, G., Akpan, I.J.: Predicting sociodemographic attributes from mobile usage patterns: Applications and privacy implications. Big Data **12**(3), 213–228 (2024)
16. Seedat, N., Jeffares, A., Imrie, F., van der Schaar, M.: Improving adaptive conformal prediction using self-supervised learning. In: International Conference on Artificial Intelligence and Statistics. pp. 10160–10177. PMLR (2023)
17. Toccaceli, P., Nouretdinov, I., Gammerman, A.: Conformal predictors for recommender systems. Machine Learning **109**(3), 643–664 (2020)
18. Toccaceli, P., Vovk, V., Nouretdinov, I.: Conformal predictors for medical diagnosis. In: Conformal and Probabilistic Prediction and Applications. pp. 165–186 (2019)
19. Tu, Z., Cao, H., Lagerspetz, E., Fan, Y., Flores, H., Tarkoma, S., Nurmi, P., Li, Y.: Demographics of mobile app usage: Long-term analysis of mobile app usage. CCF Transactions on Pervasive Computing and Interaction **3**(3), 235–252 (2021)
20. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005)