EXPLAINABILITY OF PREDICTIVE UNCERTAINTY MODELS UNDER DRIFT IN THE TELECOM DOMAIN

Nagesh Walchatwar
International Institute of Information Technology Hyderabad (IIITH)
Hyderabad, India &
Ericsson Research
Bangalore, India
nagesh.walchatwar@ericsson.com

Alberto Hata
Ericsson Research
Indaiatuba, Brazil
alberto.hata@ericsson.com

Ajay Kattepur
Ericsson Research
Bangalore, India
ajay.kattepur@ericsson.com

ABSTRACT

Machine learning models deployed in dynamic environments are prone to distributional shifts that degrade predictive reliability. While uncertainty quantification and drift detection are widely studied, their interaction with explainability remains insufficiently understood, particularly for regression tasks under covariate shift. This paper presents a unified experimental framework that integrates uncertainty quantification, calibration, drift analysis, and explainable AI (XAI) for regression models in a real-world telecommunications setting. We evaluate Bayesian neural networks (BNNs) with Monte Carlo Dropout and deep ensembles, quantifying predictive uncertainty using variance and assessing calibration quality via Expected Normalized Calibration Error. Through controlled covariate drift experiments on a real-world vehicle-to-infrastructure (V2I) communication dataset, we analyze how uncertainty degradation and calibration breakdown are reflected in explanation behavior using SHAP. The results show that BNNs exhibit higher sensitivity to drift through pronounced increases in predictive uncertainty, while ensemble models provide more stable but less adaptive estimates. Importantly, explanation patterns consistently track uncertainty degradation. A finding that has not previously been demonstrated for regression tasks under covariate drift in telecom settings, which highlights XAI as a principled diagnostic tool for drift-aware model monitoring and lifecycle management in non-stationary environments.

1 INTRODUCTION

Machine learning (ML) models deployed in real-world environments operate under the persistent challenge of non-stationary data. This is a common situation encountered in telecommunication networks, where changes in network conditions and channel characteristics induce covariate (virtual) drift (Hu et al., 2018; Oliveira et al., 2023), i.e., shifts in the input feature distribution while the underlying prediction task and label semantics remain unchanged. Such drift affects not only prediction accuracy but also the confidence a model has in its outputs. As a result, uncertainty quantification (UQ) becomes an essential capability for assessing model reliability under distributional shift in communication networks. However, beyond performance degradation, an equally important yet underexplored consequence of drift is its impact on explainable AI (XAI) methods, in particular those intended to explain predictive uncertainty models.

Recent research shows that post-hoc XAI techniques may perform differently on distribution shift (Mougan et al., 2022; Hwang et al., 2025). This suggests inconsistent explanations when the model's data distribution changes after training, limiting its application for drift detection through credit assignment variation analysis. Consequently, researchers have investigated a new paradigm towards

integrating UQ and XAI as a promising pathway to not only detect drift but also understand its underlying causes and evaluate how model confidence deteriorates or improves under adaptation (Watson et al., 2023; Mougan & Nielsen, 2023).

In this work, we investigate the interplay between uncertainty, drift detection, and explainability for regression-based performance prediction in a real-world vehicle-to-infrastructure (V2I) telecommunication dataset. We quantify predictive uncertainty under controlled virtual drift scenarios and analyze how the quantified uncertainty (specifically, epistemic uncertainty), alongside the uncertainty calibration metric, responds to these perturbations. Furthermore, we study whether XAI-driven insights can help identify drift sources and potentially reduce uncertainty through better model understanding or adaptation. Two XAI methods with different characteristics are employed: Shapley Additive Explanation (SHAP) for input feature attribution of predicted instances and Local Interpretable Model-agnostic Explanation (LIME) for sensitivity analysis of input features (Pelosi et al., 2025b). This enables verifying whether a particular XAI method is better suited to the current problem. Although XAI methods such as SHAP and LIME operate on individual predictions, their attributions depend on the underlying data distribution and feature interactions. As a result, distributional shifts can alter explanation patterns even when model architecture remains fixed, motivating the study of XAI behavior under drift.

Key Contributions As part of this in-progress research, the key contributions of this work are as follows: (i) present the first systematic experimental framework to jointly analyze predictive uncertainty, calibration behavior, and explainability under controlled covariate drift for a regression task in a real-world V2I telecom setting, which is a combination not previously studied in this domain; (ii) characterized how uncertainty estimates and calibration quality degrade as drift severity increases for Bayesian neural networks and deep ensembles, providing quantitative evidence of differential robustness between model classes; (iii) demonstrated that uncertainty-based signals serve as effective label-free indicators of distributional shift, validated across multiple drift levels; and (iv) presented that XAI methods applied to predictive uncertainty yield interpretable diagnostics that consistently and stably highlight drift-affected features both before and after calibration, establishing XAI as a reliable monitoring signal in dynamic environments.

2 Related Works

Recently, XAI has focused on interpreting challenging situations encountered by models after deployment, such as in non-stationary conditions. Most of these works investigate the effect of model drift on XAI explanations or detecting drifts through XAI. An example is the work from John (2025), who proposes adaptive explanation mechanisms that recalibrate SHAP attributions using sliding windows and drift-aware conditions for stable and informative explanations in concept drift situations. In (Edakunni et al., 2024), drifts are detected by monitoring the prediction outputs through SHAP, and identifies possible drift sources from the explanations.

Another direction followed by the literature is the adoption of UQ metrics as an explanation mechanism (Pelosi et al., 2025a). This paradigm is discussed by Deuschel et al. (2024), who argues that quantified uncertainties contribute to model trustworthiness as they can serve as a diagnostic tool to identify the source of prediction inconsistencies. This aspect is further explored for neural networks (NN) in (Thuy & Benoit, 2024b), given the overconfident nature of these models. The authors claim that UQ satisfies some XAI properties such as trust, actionability, and robustness. In particular, robustness is treated as a fundamental property when the NN faces distribution shifts.

In contrast, a few papers have studied the combination of XAI methods for predictive uncertainty models under drift conditions, which is the closest topic to the current work. Watson et al. (2023) proposes incorporating information theory metrics into the Shapley values to enhance the explanations of the entropy produced by predictive uncertainty models. The work from Mougan & Nielsen (2023) leverages bootstrapped uncertainty estimates together with SHAP explanations to monitor deployed models without labels, detecting performance drops while identifying the features most responsible for deterioration. The method constructs prediction intervals with better coverage than traditional techniques and uses explainable uncertainty to pinpoint drift-driven feature influences, enabling both detection and interpretation of distributional change. UDD triggers retraining when model uncertainty increases persistently, demonstrating superior drift detection on both regression

and classification benchmarks and illustrating how uncertainty metrics alone can be effective drift indicators. Thuy & Benoit (2024a) positions uncertainty estimation itself as an explainability mechanism for neural networks. The framework argues that uncertainty can serve as a diagnostic and trustworthy signal under drift, linking traditional XAI properties such as robustness and actionability to quantified predictive uncertainty. Löfström et al. (2025) proposes a method combining local feature importance with calibrated predictive uncertainty to generate rapid, uncertainty-aware explanations.

Despite these advances, including works on XAI-based drift monitoring (John, 2025; Edakunni et al., 2024; Pelosi et al., 2025a) and uncertainty-driven explainability (Watson et al., 2023; Mougan & Nielsen, 2023), no existing framework jointly addresses how uncertainty measurement, calibration quality, and XAI attributions of uncertainty (rather than predictions) co-evolve under controlled covariate drift in a regression setting. Most existing frameworks either focus on classification or treat UQ and XAI in isolation, without jointly considering calibration behavior and distribution shift in continuous-output models. In contrast, our work analyses how predictive uncertainty, calibration metrics, and XAI attributions (e.g., SHAP) interact across progressively increasing drift levels in a real-world telecom regression task.

3 Foundations of Predictive Uncertainty

Predictive uncertainty models usually produce the uncertainty of an output $\hat{y}$, given an input instance x. Commonly, this is represented by a probability distribution, such as Gaussian, in the form $\mathcal{N}(\mu; \sigma^2)$, where $\hat{y} = \mu$ and σ^2 is the quantified uncertainty. Depending on the adopted model or strategy to calculate the uncertainty, different types of uncertainty can be obtained (Hüllermeier & Waegeman, 2021):

- Aleatoric uncertainty: captures the inherent noise or randomness present in the data-generating process, such as measurement noise or stochastic environmental variations.
- Epistemic uncertainty: arises from a model's lack of knowledge about the underlying data distribution (Abdar et al., 2021).
- Total uncertainty: represents the summation of aleatoric and epistemic uncertainties.

This work focused on two predictive uncertainty techniques: Bayesian neural networks and deep ensembles, which primarily produce epistemic uncertainties (Blundell et al., 2015).

3.1 Uncertainty Quantification

Probabilistic models, such as Bayesian neural networks (BNNs), are a popular approach for predictive uncertainty estimation. A BNN incorporates Bayesian learning capabilities into a standard neural network by representing its weights as probability distributions (e.g., Gaussian). In practice, predictive uncertainty is approximated using Monte Carlo (MC) Dropout, where dropout layers remain active during inference, and each forward pass corresponds to a different stochastic realisation of the network induced by a randomly sampled dropout mask. Given an input $\mathbf{x}$, the predictive mean and variance (i.e., uncertainty) are computed from multiple stochastic forward passes as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}); \quad \sigma^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(f_t(\mathbf{x}) - \hat{y}\right)^2, \tag{1}$$

where $f_t(\mathbf{x})$ denotes the prediction from the t-th stochastic forward pass and T is the number of Monte Carlo samples.

Another popular method is deep ensembles, in which N deep neural networks (DNN) are used to generate a distribution of predictions. Commonly, these networks use a different seed to train the DNN to induce variability in the predictions. The predictive mean and variance are calculated in a similar way to BNN, in which N replaces the sample size T.

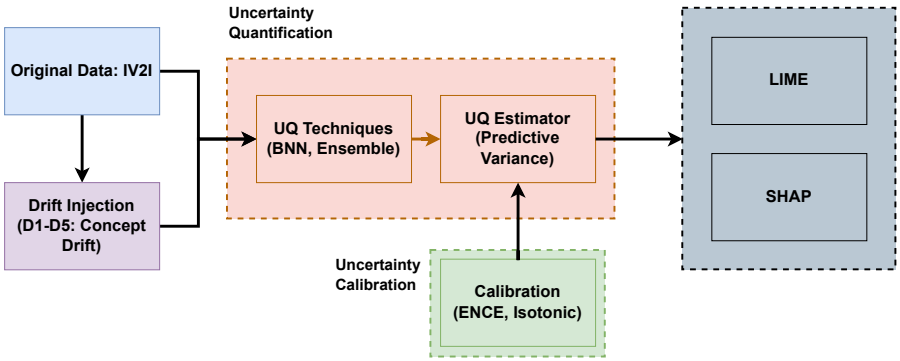


Figure 1: Proposed framework for deriving explanations from uncertainty quantification metrics by synthetically injecting concept drift on IV2I dataset from AI4Mobile.

3.2 Uncertainty Calibration and Uncertainty Calibration Metric

Commonly, the quantified uncertainties should be adjusted to reflect the prediction error magnitude. Specifically, we use isotonic regression (Niculescu-Mizil & Caruana, 2005) on a held-out calibration set to learn a non-parametric mapping from raw predictive variance to empirical squared errors. This mapping rescales uncertainty estimates without modifying model parameters and is applied to test-time predictions.

Calibration quality is evaluated using Expected Normalized Calibration Error (ENCE), which measures the discrepancy between predicted uncertainty and observed mean squared error across bins. This metric is reported both before and after calibration to evaluate the effectiveness of uncertainty correction under different levels of injected data drift (Levi et al., 2022).

4 XAI for drift detection through uncertainty quantification

A framework for covariate drift impact analysis when explaining predictive uncertainty is proposed in a network KPI prediction task. Initially, BNN and deep ensemble models are trained on an industrial vehicle to infrastructure (IV2I) dataset for uncertainty quantification and used as a baseline. Then, virtual drift is injected into the test data with different levels (0% to 25%) and used as input to the baseline models. The ENCE uncertainty calibration metric is calculated before and after each drift level to check the robustness of these models with respect to the uncertainty. Similarly, SHAP and LIME post-hoc XAI methods are applied to the predicted uncertainties, before and after calibration, for the different drift levels. An overview of this process is presented in Figure 1 and the details are presented in the subsections.

4.1 XAI Integration

Explainable Artificial Intelligence (XAI) techniques are employed to interpret model behaviour under distribution drift by providing insight into predictive uncertainty variations. While uncertainty- and drift-based signals indicate when model reliability degrades, XAI is used to identify which input features contribute to changes in predictions and associated uncertainty (i.e., post-hoc local explanation). The following subsections describe the two feature attribution methods adopted in this work and their integration for explaining uncertainty behaviour and drift effects.

4.1.1 SHAP-Based Feature Attribution

SHAP is employed to obtain consistent feature-attribution scores that quantify the contribution of each input feature to a model's prediction (Lundberg & Lee, 2017). For a given input instance $\mathbf{x}$ and model prediction $\hat{y}$, SHAP computes feature attributions.

In the experiments, SHAP values are computed for samples within both reference and drifted time windows. Let $\Phi_t = \{\phi_i^{(j)}\}$ denote the set of SHAP values for samples j in time window t. Changes in feature influence are quantified by comparing the mean absolute SHAP values across windows:

$$\Delta_{\text{SHAP}}(i,t) = \left| \mathbb{E}_{j \in t} \left[|\phi_i^{(j)}| \right] - \mathbb{E}_{j \in \text{ref}} \left[|\phi_i^{(j)}| \right] \right|. \tag{2}$$

Features with large $\Delta_{\text{SHAP}}(i,t)$ are identified as primary contributors to drift- or uncertainty-induced prediction changes.

4.1.2 LIME-Based Local Explanations

LIME is used to generate instance-level explanations for samples exhibiting high predictive uncertainty or anomalous behavior (Ribeiro et al., 2016). Given an input instance $\mathbf{x}$, LIME constructs a local surrogate model $g(\mathbf{x})$ by minimizing

$$\mathcal{L}(f, g, \pi_{\mathbf{x}}) + \Omega(g), \tag{3}$$

where f is the original black-box model, $\pi_{\mathbf{x}}$ is a proximity kernel that weights perturbed samples according to their similarity to $\mathbf{x}$, and $\Omega(g)$ is a regularization term enforcing interpretability.

In the experimental setup, LIME explanations are generated for selected samples within drifted windows, particularly those with high standardized residuals or elevated uncertainty.

4.1.3 XAI Integration with Uncertainty and Drift Signals

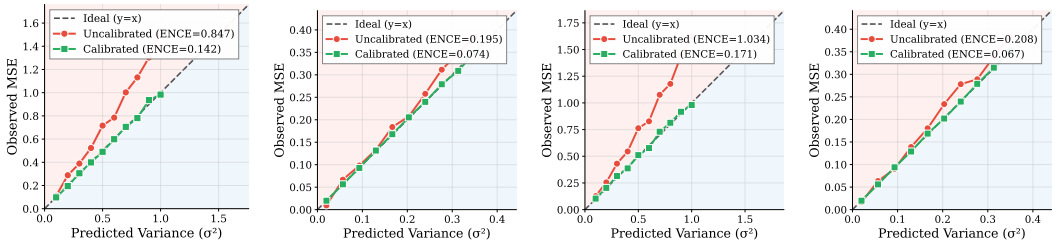
SHAP and LIME explanations are jointly analyzed alongside uncertainty and calibration metrics. By correlating feature-attribution changes with increases in predictive uncertainty or ENCE, the framework enables interpretable drift diagnosis rather than simple drift detection. This integrated analysis supports informed adaptation strategies, such as targeted retraining or feature-level monitoring, while maintaining transparency and accountability in dynamic environments. The objective is not to detect drift via SHAP values, but to interpret how uncertainty increases are attributed to specific features once drift has already been identified through uncertainty signals.

5 Experiments and Results

The experimental pipeline integrates UQ, controlled drift injection, calibration assessment, drift detection, and explainability-driven analysis to systematically evaluate the behavior of uncertainty-aware regression models under covariate drift (overall diagram in Figure 1). The objective of this experimental setting is to analyze how predictive uncertainty evolves under increasing drift severity, how calibration affects uncertainty reliability, and how explainability methods can be used to diagnose and mitigate uncertainty degradation.

5.1 Experimental Setting

All experiments are conducted on the *AI4Mobile Industrial Wireless Dataset* (Hernangomez et al., 2022). Specifically, the IV2I+ dataset is used, which contains communication KPIs from an industrial mobile robot communicating with a base station. This dataset exhibits strong temporal and



(a) BNN (0% drift) (b) Ensemble (0% drift) (c) BNN (25% drift) (d) Ensemble (25% drift)

Figure 2: Comparison of uncertainty calibration performance for BNN and ensemble models under baseline and 25% covariate drift conditions.

spatial dynamics, as evidenced by variations in network KPIs, mobility patterns, and robot sensor data. Such variations naturally introduce distributional changes driven by evolving network load, robot motion trajectories, or environmental factors.

Standard preprocessing steps are applied, including missing-value imputation, feature normalization, and removal of non-predictive identifiers. All numerical features, including radio KPIs such as RSRP, RSRQ, RSSI, SNR, and mobility/sensor readings, are retained as model inputs X. The regression target Y is uplink throughput (in Mbps), which represents the data rate achieved by the mobile robot communicating with the base station. Predicting throughput from radio and context features is a standard network performance estimation task. Drift is injected exclusively into a subset of input features (RSRP, RSRQ, RSSI, SNR) in the test partition, thereby simulating covariate shift while the semantics of the target variable remain unchanged. The dataset is split into training and testing partitions, with drift injected exclusively into the test data to preserve a clean training distribution.

We train two classes of uncertainty-aware regression models using only non-drifted training data. First, a *deep ensemble* is constructed using five multilayer perceptron (MLP) regressors, each composed of three hidden layers with 128, 64, and 32 neurons and ReLU activations. Ensemble members are trained with different random initializations and bootstrap-resampled versions of the training data to encourage model diversity. While MC Dropout and deep ensembles primarily capture epistemic uncertainty, we treat predictive variance as a practical proxy for total uncertainty, following common practice in uncertainty-based monitoring under distribution shift. Predictive uncertainty for the ensemble is quantified as the variance across ensemble member predictions. Second, a *Bayesian Neural Network (BNN)* is implemented using the same three-layer MLP architecture, with dropout layers (dropout rate $= 0.2$) retained during inference. Predictive uncertainty is estimated via Monte Carlo Dropout to approximate Bayesian posterior sampling. Both models share the same base architecture and serve as baseline predictors for all subsequent uncertainty, drift, and explainability analyses.

We apply a standard post-hoc uncertainty calibration through isotonic regression to align predicted uncertainty with observed prediction errors. For the clean test set (baseline condition), both models produce predictive means along with total uncertainty, which in this work is quantified exclusively using variance. For BNN, variance is obtained from the dispersion of Monte Carlo predictive samples, while for ensemble models, it is computed from the variability across ensemble members (subsection 3.1).

A controlled synthetic drift setup is deliberately chosen here as a validation strategy: by knowing exactly which features are perturbed and by how much, we can verify whether SHAP and LIME correctly recover the ground-truth drift sources. This serves as a necessary sanity check before claiming diagnostic value in uncontrolled settings. To simulate controlled virtual drift, synthetic perturbations are injected into a subset of high-impact input features (RSRP, RSRQ, RSSI, and SNR) while preserving class labels. Drift is introduced by uniformly selecting a fraction $p \in \{5\%, 10\%, 15\%, 20\%, 25\%\}$ of test samples. For each selected sample and feature f, the original feature value is shifted by an additive offset proportional to the feature's empirical standard deviation σ_f, such that

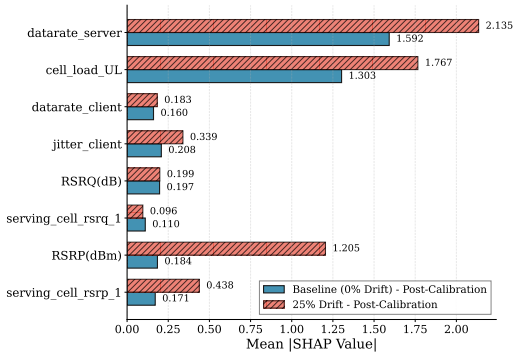
$$x_f^{\text{drift}} = x_f + \alpha \cdot \sigma_f,$$

where α controls the drift magnitude (e.g., 10% drift level indicates that 10% of the test samples are perturbed). This procedure induces progressively stronger distributional shifts without altering label semantics. All drifted sample indices are logged to enable direct comparison between clean and drifted instances and precise evaluation of drift detectability. Feature-based detectors capture input distribution changes, whereas uncertainty-based signals reflect model confidence degradation. Our focus is on the latter, as it is model-agnostic and directly tied to reliability.

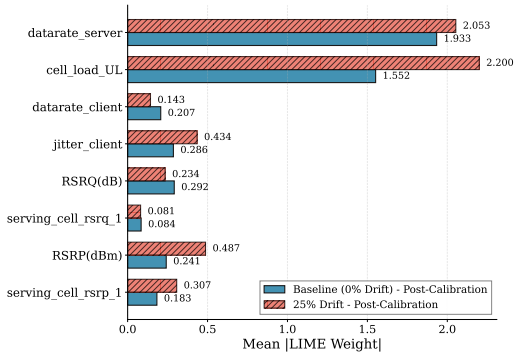
To assess the reliability of predicted uncertainty, we evaluate calibration using ENCE. Isotonic calibration is applied to both models. Table 1 presents uncalibrated and calibrated ENCE values across different drift levels, enabling a direct comparison of calibration robustness under distribution shift.

Table 1: Uncalibrated and calibrated ENCE comparison for BNN and deep ensemble. Lower ENCE values resemble better uncertainties produced by the model.

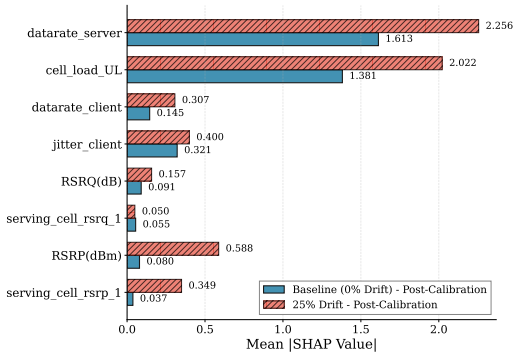
	BNN ENCE		Deep Ensemble ENCE	
Drift Level	Uncalibrated	Calibrated	Uncalibrated	Calibrated
0% (baseline)	0.847	0.142	0.195	0.074
10%	0.912	0.158	0.199	0.069
25%	1.034	0.171	0.193	0.070
50%	1.256	0.203	0.208	0.067



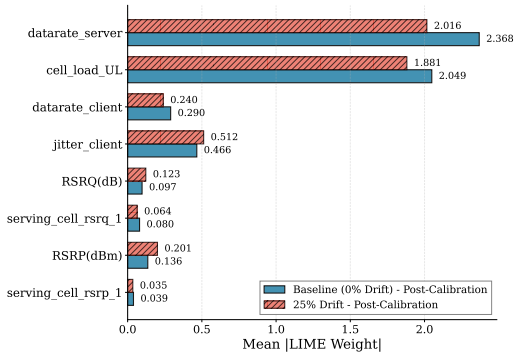
(a) BNN (0% & 25% drift) – SHAP



(b) BNN (0% & 25% drift) – LIME



(c) Ensemble (0% & 25% drift) – SHAP



(d) Ensemble (0% & 25% drift) – LIME

Figure 3: XAI analysis of predictive uncertainty (post-calibration) under baseline (no drift) and 25% covariate drift for BNN and ensemble models. SHAP provides global feature-level attribution shifts, while LIME offers complementary local explanations.

5.2 Experimental Results

5.2.1 Uncertainty and Calibration Behaviour under Covariate Drift

We first examine how predictive uncertainty behaves as covariate drift severity increases. For both BNN and ensemble models, uncertainty-related quantities exhibit a clear increasing trend with drift. As shown in Table 1, the uncalibrated uncertainty estimates for BNNs increase substantially from the baseline to higher drift levels, reflecting growing epistemic uncertainty as the test distribution diverges from the training data. Ensemble models also show an increase in uncertainty under drift, although the magnitude of change is smaller and more gradual, indicating comparatively smoother and more stable behaviour.

We next evaluate the effect of post-hoc calibration using ence. For the baseline (0% drift) case, calibration consistently improves uncertainty quality for both model classes, reducing ence from 0.847 to 0.142 for BNNs and from 0.195 to 0.074 for ensembles. Under increasing drift, however, calibrated ence values rise steadily for both models, indicating degradation in calibration quality as distribution shift intensifies. This effect is particularly pronounced for BNNs, where calibrated ence

increases from 0.142 at baseline to 0.203 at 50% drift. Ensemble models exhibit a similar but milder trend, with calibrated ence remaining lower overall but still increasing under drift. These results demonstrate that calibration learned on in-distribution data does not remain reliable under covariate shift, highlighting the brittleness of static post-hoc calibration methods in drifting environments. The calibration step incurs minimal computational overhead, as isotonic regression is trained offline on a small calibration set and introduces only a constant-time mapping during inference.

We analyze how uncertainty estimates evolve under increasing levels of covariate drift and assess their suitability as label-free drift indicators. Predictive uncertainty measures are aggregated over sliding windows and tracked across progressively stronger drift conditions. As drift severity increases, all uncertainty signals exhibit a systematic upward trend, indicating reduced model confidence even before substantial drops in predictive accuracy are observed.

5.2.2 XAI Behaviour under Drift

To analyze how model explanations evolve under covariate drift, we apply SHAP and LIME to the uncertainty outputs rather than to point predictions. Global SHAP analysis shows a systematic reallocation of feature attributions as drift severity increases, with features explicitly perturbed during drift injection exhibiting consistently higher contributions to predictive uncertainty. At the instance level, LIME explanations of samples with elevated uncertainty confirm that these uncertainty increases are primarily driven by the same drift-affected features. Figure 3 shows SHAP and LIME results after applying different levels of drift to the predictive uncertainty models.

5.2.3 Impact of Calibration on XAI Attributions

Fig. 3 illustrates the influence of calibration on explainability by comparing SHAP and LIME attributions for BNN and ensemble models under baseline (0%) and drifted (25%) conditions. Across all settings, calibration reduces the overall magnitude of uncertainty-driven attributions, reflecting the compression of predictive variance after calibration. However, the relative ordering of dominant features remains largely unchanged: features directly affected by drift injection (e.g., RSRP, RSRQ, RSSI, and SNR) consistently retain high importance both before and after calibration. This stability is observed for both global SHAP attributions and local LIME explanations, indicating that calibration primarily rescales attribution strength without altering feature relevance. These results demonstrate that calibrated uncertainty preserves the semantic consistency of XAI outputs, supporting the robustness of explainability analyses under calibrated predictive uncertainty.

5.2.4 Summary of Findings

Overall, the results indicate that (i) predictive uncertainty exhibits a consistent and monotonic increase as covariate drift severity grows; (ii) calibration improves the reliability of uncertainty estimates primarily under in-distribution conditions and degrades under stronger drift; (iii) uncertainty-based signals provide effective label-free indicators of distributional shift; and (iv) explainability methods applied to uncertainty estimates yield interpretable diagnostics that correctly highlight drift-affected features.

6 Conclusion and Future Directions

This paper presented an in-progress experimental study on the interaction between predictive uncertainty, calibration behaviour, drift detection, and explainability under controlled covariate (virtual) drift in a real-world telecom regression setting. Through empirical analysis of Bayesian neural networks and deep ensembles, we showed that uncertainty estimates and calibration quality degrade systematically as drift severity increases, while uncertainty-based signals remain effective for label-free drift indication. Furthermore, we demonstrated that explainability methods applied to predictive uncertainty provide stable and interpretable diagnostics that consistently identify drift-affected features, both before and after calibration. As part of ongoing work, we evaluate XAI-on-uncertainty under natural drift using temporal splits of the IV2I+ dataset and additional telecom datasets. We also explore adaptive uncertainty calibration and explanation models that evolve under continuous drift, and analyze the impact of concept drift on explanation fidelity using different predictive uncertainty models.

ACKNOWLEDGMENT

This work is supported by the *Horizon Europe project PANDORA* under grant agreement number 101135775.

REFERENCES

Moloud Abdar et al. A review of uncertainty quantification in deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(12):5415–5437, 2021.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1613–1622. JMLR.org, 2015.

Jessica Deuschel et al. The Role of Uncertainty Quantification for Trustworthy AI. In *Unlocking Artificial Intelligence*, pp. 95–115. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-64831-1 978-3-031-64832-8. doi: 10.1007/978-3-031-64832-8_5. URL https://link.springer.com/10.1007/978-3-031-64832-8_5.

Narayanan U. Edakunni, Utkarsh Tekriwal, and Anukriti Jain. Explaining drift using shapley values, 2024. URL https://arxiv.org/abs/2401.09756.

Rodrigo Hernangomez et al. AI4Mobile Industrial Wireless Datasets: iV2V and iV2I+, 2022. URL https://dx.doi.org/10.21227/04ta-v128.

Hanqing Hu, Mehmed Kantardzic, and Lingyu Lyu. Detecting Different Types of Concept Drifts with Ensemble Framework. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 344–350, 2018. doi: 10.1109/ICMLA.2018.00058.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Hyunseung Hwang et al. SHAP-based Explanations are Sensitive to Feature Representation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1588–1601, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732105. URL https://doi.org/10.1145/3715275.3732105.

Shivogo John. Fair and Explainable Credit-Scoring under Concept Drift: Adaptive Explanation Frameworks for Evolving Populations. *arXiv preprint arXiv:2511.03807*, 2025. URL https://arxiv.org/abs/2511.03807.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and Calibrating Uncertainty Prediction in Regression Tasks. *Sensors*, 22(15), 2022. ISSN 1424-8220. doi: 10.3390/s22155540. URL https://arxiv.org/abs/2410.21129.

Johan Löfström et al. Efficient and uncertainty-aware explanations for machine learning models. *arXiv preprint arXiv:2509.XXXX*, 2025.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Carlos Mougan, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Explanation Shift: Detecting distribution shifts on tabular data via the explanation space. 2022.

Charles Mougan and Didrik Nielsen. Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap. *arXiv preprint arXiv:2201.11676*, 2023.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting Good Probabilities With Supervised Learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005.

Gustavo H. F. M. Oliveira et al. Tackling Virtual and Real Concept Drifts: An Adaptive Gaussian Mixture Model Approach. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):2048–2060, 2023. doi: 10.1109/TKDE.2021.3099690.

D. Pelosi et al. Explainability and interpretability in concept and data drift. *Algorithms*, 18(7):443, 2025a. URL https://www.mdpi.com/1999-4893/18/7/443.

Daniele Pelosi, Diletta Cacciagrano, and Marco Piangerelli. Explainability and Interpretability in Concept and Data Drift: A Systematic Literature Review. *Algorithms*, 18(7):443, July 2025b. ISSN 1999-4893. doi: 10.3390/a18070443. URL https://www.mdpi.com/1999-4893/18/7/443.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Arthur Thuy and Dries Benoit. Explainability through Uncertainty: Trustworthy Decision-Making with Neural Networks. *arXiv preprint arXiv:2403.10168*, 2024a.

Arthur Thuy and Dries F. Benoit. Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, 317(2):330–340, September 2024b. ISSN 03772217. doi: 10.1016/j.ejor.2023.09.009. URL http://arxiv.org/abs/2403.10168.

David S. Watson et al. Explaining Predictive Uncertainty with Information Theoretic Shapley Values, oct 2023. URL http://arxiv.org/abs/2306.05724. https://arxiv.org/abs/2306.05724.