Consistency of Explanations in Evolving Supervised Learning Contexts

Marcin Ostrowski[1[0009−0006−5845−2976]], Katarzyna Kaczmarek-Majer[1,2[0000−0003−0422−9366]], and Jose M. Alonso-Moral[3[0000−0003−3673−421X]]

[1] Systems Research Institute Polish Academy of Sciences,
Newelska 6, 01-147 Warsaw, Poland
{Marcin.Ostrowski,k.kaczmarek}@ibspan.waw.pl

[2] Institute for Research and Applications of Fuzzy Modeling, University of Ostrava,
Ostrava, Czech Republic

[3] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
josemaria.alonso.moral@usc.es

Abstract. In many real-world scenarios, explanations are not perceived as isolated entities, but rather as streams accompanying models that operate over extended periods of time. In such contexts, users are exposed not to a single explanation, but to a sequence of explanations whose evolution may influence trust, understanding, and decision-making. In this work, we aim to capture the evolution of explanations that exhibit sudden changes, contradict prior explanations, or demonstrate erratic fluctuations, as these may compromise their practical utility, even when maintaining a high level of predictive performance. We propose a framework aimed at supporting the assessment of the consistency of explanations for evolving supervised learning contexts, taking into account the variability of the input datasets, and the uncertainty of the model. We introduce a model-agnostic and explanation-agnostic index for quantifying the temporal consistency of automated explanations. We empirically demonstrate that the proposed index facilitates the quantification and localization of temporal instability in explanation streams. Experimental results for simulated multivariate time series underscores the ability of the proposed index to capture instantaneous changes and their relative impact in relation to historical patterns.

Keywords: Explainable Artificial Intelligence · Online Learning · Stability · Classification · Fuzzy Linguistic Summarization

1 Introduction

This work is accepted for publication in Proceedings of the International Conference

Explanations are the foundation of trustworthy Artificial Intelligence (AI) systems, playing a central role in improving transparency and user confidence across various decision-support systems [1,6,8,10,18]. The Rashomon effect was first introduced by Breiman [3] and then made popular in the context of Explainable AI (XAI) by Rudin [20]. This effect justifies why sometimes different models explain the data-generating process equally well, but they may contradict each other because their predictions are built in a different way, i.e., following a different rational (e.g., relying on different features). Accordingly, in case that multiple, equally good models exist, their interpretations should be carefully compared before drawing any conclusion. If this is the case, selecting one specific model for a particular task should depend on multiple factors (e.g., performance, interpretability, explainability, fairness, sparsity, stability or consistency).

In this work, we focus on real-world scenarios, when observations are accumulated over time, often in batches, reflecting the evolution of the environment and the transformation of data-generating processes [7,22]. We observe that despite the growing body of research on XAI, most existing approaches implicitly assume static tabular data or static data distributions. The field of temporal scenarios has traditionally focused on generating explanations within designated sliding or expanding time windows [19]. This process commonly involves applying smoothing or aggregation techniques to static explanation methods in the temporal domain [11,21]. While these approaches facilitate temporal attribution analysis, they offer limited insight into the consistency of explanations themselves. Research areas that are closely related, such as statistical process control and data distribution drift detection, study temporal changes in data or model outputs [2, 7]. However, these areas are not designed to assess explanations as structured, semantic objects. Another direction of research explores explainable online classification approaches, e.g., using prototypes [4]. In this research, we further extend previous work by proposing a pipeline for monitoring the consistency of explanations in evolving supervised learning contexts.

In this study, we posit that the temporal consistency of explanations constitutes a complementary dimension for model evaluation alongside predictive performance. Our research question focuses on developing and validating a convenient measure to capture how explanations change over time, accounting for their history and the various assumptions underlying both predictive modeling and the procedure used to derive explanations. The analysis of this question facilitates comparisons not only between explanations within a single model, but also between different models that may demonstrate similar performance, yet yield explanations of differing temporal reliability. To summarize, the main

on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU2026).

Please cite this article in press as:

Ostrowski, M., Kaczmarek-Majer, K., Alonso-Moral, J.M. (2026). Consistency of Explanations in Evolving Supervised Learning Contexts. In: Vantaggi, B., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2026. Communications in Computer and Information Science, vol 3019. Springer, Cham. https://doi.org/10.1007/978-3-032-28994-0_36

contribution of this work is the introduction of a novel index for measuring the temporal consistency of automated explanations as the quantification of the deviations from the expected temporal behavior of explanation signals. The proposed index is model-agnostic and explanation-agnostic, relying solely on the temporal progression of explanatory quantities. We demonstrate through a series of illustrative experiments that temporal consistency can reveal significant information about explanation reliability. In particular, it is demonstrated that models with comparable accuracy may exhibit substantial discrepancies in the temporal stability of their explanations. This underscores the significance of explanation consistency as a criterion for model selection and implementation in evolving environments.

The rest of the manuscript is organized as follows. Section 2 introduces the new index. Section 3 describes the experimental settings. Section 4 presents the main reported results. Section 5 provides readers with concluding remarks and outlines future work.

2 Temporal Consistency Assessment of Explanations

Figure 1 provides a schematic overview of the proposed temporal consistency assessment pipeline adapted to a showcase example. The aim of this pipeline is to provide a quantitative measure of explanation consistency based on the input data. The design is intended to be both model- and explanation-agnostic. The proposed approach consists of three main sequential steps: (i) predictive modeling; (ii) explanation generation; and (iii) consistency monitoring.

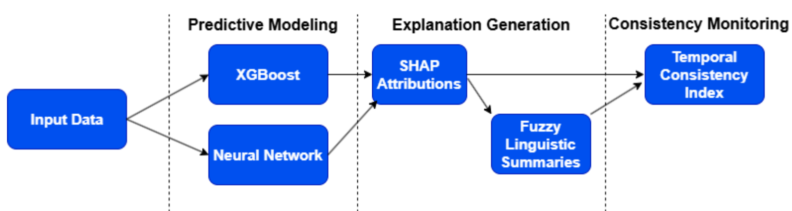

Fig. 1: Overview of the temporal consistency assessment pipeline adapted to an example. Explanatory signals generated by a predictive model over successive time windows are treated as temporal processes and evaluated using the proposed index, yielding a quantitative measure of explanation consistency.

2.1 Predictive Modeling

The field of predictive modeling in evolving environments has seen significant attention in the context of concept drift and online learning. The overarching aim is to preserve the efficacy of predictive models in environments where data distributions are constantly fluctuating [7, 22]. In the domain of XAI, predictive models are commonly regarded as fixed structures, with explanations automatically generated after the learning process and examined independently from the underlying learning mechanics [8, 11]. However, comparatively little attention has been paid to how the selection and development of the predictive model itself influences the temporal behavior of explanations.

In this study, predictive models are utilized as sources of explanatory signals, rather than as subjects of optimization. As a result, the focus is intentionally limited to well-established supervised learning models. Specifically, we consider gradient-boosted decision trees implemented via XGBoost [5] and a simple neural network represented by a multilayer perceptron classifier from the scikit-learn [17]. The present work does not seek to provide a comprehensive comparison of predictive models. Instead, it focuses on assessing explanation consistency under representative modeling choices.

2.2 Automated Explanation Generation

Contemporary explanation techniques, encompassing feature attribution and rule-based methodologies, are predominantly designed to provide local explanations for individual predictions or global explanations for static models [14, 18]. In the context of temporal data analysis, explanations are typically generated independently at each temporal interval and subsequently integrated or smoothed for visualization [11].

In this work, the features' attributions derived from SHAP framework [11] are regarded as time-indexed explanatory signals. For a given feature j, Shapley values are computed for individual observations at a specific time step. In addition to numerical attribution streams, we construct fuzzy linguistic summaries (FLS) [9] that encapsulate global, human-interpretable statements regarding the behavior of explanations. These summaries are formulated using extended protoforms [13, 23], where the relation between SHAP-derived quantities and feature values serves as the underlying numerical evidence. In case of having big datasets with many attributes, the so-called SHAP length metric can be used to select the most informative attributes for the desired degree of explanation completeness [12]. The proposed framework is explanation-agnostic and operates on arbitrary scalar explanatory streams. The present work focuses on SHAP-based attributions and fuzzy linguistic summaries.

2.3 Temporal Consistency Monitoring

Let $F = \{F_t\}_{t=1}^{T}, T > 1$, denote a real-valued temporal process describing the evolution of an explanatory quantity over discrete time steps, where $F_t \in \mathbb{R}$ summarizes the explanatory information available at time t, for example via aggregation over a batch of observations. Our objective is to quantify the temporal consistency of the process F, understood as the extent to which its evolution can be regarded as stable and predictable over time, as opposed to exhibiting abrupt or erratic deviations.

To assess deviations from expected temporal behavior, we construct a one-step predictor $\hat{F}_t$ of F_t based on past values. In this work, we employ a modified exponentially weighted moving average (EWMA) with a drift correction term, allowing the predictor to adapt to smooth trends while remaining sensitive to gradual changes in the underlying process

$$\hat{F}_t = (1-\lambda)\hat{F}_{t-1} + \lambda F_{t-1} + \beta d_{t-1}, \qquad d_t = (1-\mu)d_{t-1} + \mu(F_t - F_{t-1}),$$

with initialization $\hat{F}_1 = F_1$ and $d_1 = 0$.

The parameter $\lambda \in [0,1]$ governs the degree to which the one-step predictor $\hat{F}_t$ adapts to the most recent observed value $\hat{F}_{t-1}$. Setting λ to a higher value, results in the predictor reacting in a stronger way to recent changes in the observed explanatory signal. This can be interpreted as the predictor becoming more responsive to short-term fluctuations. Conversely, setting λ to a smaller value, yields a smoother and more inertial baseline due to the predictor relying more heavily on its previous estimate $\hat{F}_{t-1}$.

The parameter $\beta \in [0,1]$ determines the influence of the estimated drift term d_{t-1} on the predictor $\hat{F}_t$. With larger values of β, the predictor incorporates more information about the recent trend of the process. This improves the ability of the predictor to follow gradual systematic changes. When the parameter β is smaller, the predictor behaves more like a standard exponentially weighted moving average without explicit trend adjustment.

The parameter $\mu \in [0,1]$ controls the update rate of the drift term d_t. This term summarizes the recent tendency of the signal to increase or decrease. Larger values of μ assign greater importance to the most recent difference $F_t - F_{t-1}$, making the drift estimate more sensitive to newly emerging trends. Smaller values of μ produce a more stable and slowly varying drift estimate, reflecting longer-term directional behavior rather than short-lived changes.

Deviations from expected behavior are quantified using normalized residuals. Let

$$s = \mathrm{MAD}\big(\{F_t\}_{t=1}^{T}\big)\,,$$

denote a robust scale estimator of the process based on the median absolute deviation (MAD). The normalized residuals are defined as

$$r_t = \frac{F_t - \hat{F}_t}{s}, \quad t = 2, \ldots, T.$$

If $s = 0$, indicating a constant process, we define the temporal consistency as equal to its maximum value.

The Temporal Consistency Index (TCI) of the process F is defined as

$$\mathrm{TCI}(F) = 1 - \sum_{t=2}^{T} \omega_t\, \ell(r_t)\,, \quad \ell(r) = \frac{r^2}{\tau^2 + r^2}, \tau > 0.$$

where $\omega_t \geq 0$, $\sum_{t=2}^{T} \omega_t = 1$, and the parameter τ controls the sensitivity of the loss function to normalized residual magnitude. When τ is set to a smaller value, even moderate changes produce relatively large loss values, making the TCI more sensitive to deviations from predicted behavior. Conversely, applying large values of τ results in the loss growing more slowly, leading to only substantial deviations reducing the overall consistency.

In this work, a bounded quadratic-type loss is used to smoothly interpolate between quadratic penalization for small deviations and saturation for large deviations. By default, uniform weights $\omega_t = \frac{1}{T-1}$ are used, ensuring that all time points contribute equally to the consistency assessment. In addition to the overall index, the above formulation inherently incorporates a per-step inconsistency contribution. For each time step $t = 2, \ldots, T$ we define

$$a_t := \omega_t \, \ell(r_t)$$

which quantifies the contribution of time step t to the overall inconsistency of the explanatory signal. By construction, $a_t \geq 0$. The quantities a_t thus facilitate the identification of time points at which the explanation stream exhibits the most significant deviation from its anticipated temporal behavior.

TCI ranges from 0 to 1, with higher values denoting greater stability and predictability in temporal behavior. Conversely, lower values indicate increased variability or abrupt changes over time. It is imperative to note that the index is model-agnostic and explanation-agnostic, contingent solely on the temporal progression of the explanatory signal. In scenarios where multiple explanatory signals are available, the index is computed independently for each signal. This results in a set of consistency scores that can be analyzed collectively.

3 Experimental Settings

3.1 About Data and Evaluation Protocol

We illustrate our framework with temporal consistency assessment via a simulation experiment. Let us consider controlled non-stationarity, a synthetic multivariate time series generator is employed to produce three explanatory variables, denoted by x_1, x_2, x_3 respectively, and a binary label y taking value 0 or 1 to indicate stable versus unstable regimes respectively. The stream is divided into successive windows of equal length. Stable windows are characterized by smooth oscillatory dynamics with mild modulation in amplitude, frequency and noise. In contrast, unstable windows give rise to distributional changes, including abrupt mean shifts, gradual mean drifts, and increased noise with interaction effects.

Fig. 2 depicts exemplary simulated time series with these three types of drift. Thus, we split the sequence into twenty four segments of equal length. Indeed, six quasi-periodic segments are repeated four times.

For the first, the third and the fifth segments, we consider the following three time series: $x_1(t) = \sin(0.05t) + \varepsilon_1(t)$; $x_2(t) = \cos(0.05t) + \varepsilon_2(t)$; $x_3(t) = \sin\big(0.05t + \frac{\pi}{4}\big) + \varepsilon_3(t)$ where $\varepsilon_i(t) \sim \mathcal{N}(0,\ 1)$. Then, in the second segment, we shift each variable by a constant (mean shift) of $c_1 = 2, c_2 = -2$ and $c_3 = 1.5$. Then, for the fourth segment, incremental drift is introduced. For the sixth

Fig. 2: Simulated time series data with stable and unstable regimes.

segment, the generation mechanism changes, and the noise is particularly bigger: $x_1(t) = 0.01\,t + \eta_1(t)$; $x_2(t) = 0.8\,x_1(t) + \eta_2(t)$; $x_3(t) = \cos(0.03\,t) + \eta_3(t)$, where $\eta_i(t) \sim \mathcal{N}(0,\ 0.4^2), i = 1, 2, 3$.

The distribution generation process was then repeated every second segment for the stable class and every sixth segment for the unstable class with parameter t adapted to the place of segments in the data stream.

The temporal protocol we employ is of a batch-style nature. A predictive model is trained on initial prefix of the stream in order to include all potential forms of instability. This training set consisted of six first segments of the time series. The trained model is subjected to an evaluation process that utilizes a sequence of non-overlapping test blocks, with each test block comprising a pair of consecutive segments (one stable and one unstable) that are of equal size. This configuration emulates deployment monitoring, whereby a model that has been trained on historical data is evaluated as new batches are received over time.

Subsequent to this procedure, the test data yielded by the process comprised a total of nine test windows, each of which was subdivided into two segments.

3.2 Predictive Modeling and Explanation Generation

In accordance with the evaluation protocol described in the preceding subsection, the predictive models were subjected to independent assessment on each test window, employing the area under the ROC curve as the quality metric. Each test window corresponds to newly arriving data that was not used during training at the given time step. Table 1 summarizes the metric values across successive evaluation windows.

As expected, given the controlled simulation setup, the predictive performance remains consistently high across all test windows. Both models yield AUC close to 1 with the lowest value being 0.93 for the neural network in the time window $9-10$. It is important to note that this behavior is intentional. The primary objective of the conducted experiments is not to optimize the predictive accuracy. Instead, the focus is on the isolation and analysis of the temporal behavior of explanations under stable predictive performance. Consequently, observed variations in explanatory signals cannot be attributed to deteriorating model quality.

Table 1: AUC results for XGBoost (XGB) and neural network (NN) models evaluated over different time windows.

Model	$7-8$	$9-10$	$11-12$	$13-14$	$15-16$	$17-18$	$19-20$	$21-22$	$23-24$
XGB	1.0	0.95	0.98	1.0	0.96	0.99	1.0	0.95	0.99
NN	1.0	0.93	0.98	1.0	0.97	0.98	1.0	0.97	0.99

For each designated testing period, explanations were computed in the form of SHAP feature attributions, which were applied to all test observations. The local attributions were subsequently aggregated to form time-indexed explanatory signals, which were used as inputs to the proposed temporal consistency analysis. Specifically, for each feature, the distribution of Shapley values within a test window was summarized into a single representative value, yielding a sequence of explanation values evolving over time.

3.3 Temporal Consistency Monitoring

In order to obtain a scalar explanation signal suitable for temporal consistency assessment, the SHAP attributions calculated according to [11] are aggregated across observations in the block via median for each feature j

$$F_t^{(j)} = \operatorname{median}_{i=1,\dots,m_t}\big(\phi_t^{(i,j)}\big),$$

where $\phi_t^{(i,j)}$ denotes the SHAP attribution for observation i in block t and m_t is the number of observations in that block. The resulting sequence $\{F_t^{(j)}\}$ forms a temporal explanatory signal, for which the temporal consistency index is computed $\mathrm{TCI}(\{F_t^{(j)}\})$ independently. In addition to the calculation of the overall index, the per-step inconsistency contributions are reported to localize the time points at which the explanation stream deviates most strongly from its expected temporal behavior.

4 Experimental Results

4.1 SHAP-based Explanation Streams

The SHAP-based results for feature x1 are presented in Fig. 3, including: (i) boxplots of SHAP attribution distributions across successive test windows, (ii) the corresponding aggregated explanatory signal $F_t^{(1)}$ obtained via the median, and (iii) the per-step inconsistency contributions derived from the proposed temporal consistency index. Additional plots containing results for neural network models are available at [15].

The analysis is conducted separately for the stable and unstable classes. This enables a comparative investigation of explanation dynamics under different data-generating regimes. For the stable class, the SHAP attribution distributions and the resulting aggregated explanatory signal demonstrate notable stability over time, with only minor deviations observed in the final test windows. These deviations are reflected by elevations in the per-step inconsistency contribution. Such elevations indicate localized deviations from the expected temporal behavior. This behavior is consistent with the fact that the stable class is generated from an approximately stationary distribution.

In contrast, the explanatory signal demonstrates notable temporal variation for the unstable class, characterized by observations drawn from multiple distributions that exhibit recurrence over time. This variation results in significantly elevated per-step inconsistency contributions, which manifest in a periodic pattern synchronized with the recurrent distributional changes. In particular, windows corresponding to distributional transitions (e.g., $11-12$, $17-18$, and $23-24$) are consistently identified as major contributors to explanation inconsistency. It is worth noting that even when the explanatory signal returns to a regime similar to earlier windows, elevated per-step inconsistency may persist due to the temporal dependence embedded in the consistency computation.

This finding underscores the ability of the proposed index to not only capture instantaneous changes but also their relative impact in relation to recent historical patterns. The findings indicate that the framework facilitates the quantification and localization of temporal instability in explanation streams. The method does not intend to explain the causal origin of such instabilities; rather, it aims to identify their occurrence and magnitude for subsequent analysis.

4.2 Fuzzy Linguistic Summary-based Explanation Streams

To further assess the proposed framework, experiments were conducted using FLS as an alternative form of explanation. We briefly recall that a fuzzy linguistic summary takes the form

$$S : Q\ R_1 \star \ldots \star R_k\ x \text{ are } P_1 \diamond \ldots \diamond P_l, \tag{1}$$

where $\star, \diamond \in \{\text{AND, OR}\}$, Q is a linguistic quantifier, $R_1, \ldots, R_k$ are qualifiers and $P_1, \ldots, P_l$ are summaries.

In this setting, explanatory streams were formulated using the product of the degree of truth and the degree of support associated with each summary. The derived summaries were based on feature values extracted from the test data and their corresponding SHAP attributions for the unstable class. The fuzzy membership functions for each feature were defined as triangular fuzzy numbers constructed based on data-driven quartiles. We considered the empirical first quartile, median, and third quartile. The terms corresponding to the first and third quartile were modeled as left- and right-shoulder membership functions which decrease or increase linearly from median accordingly. An example for feature x1 is showcased in Fig. 4.

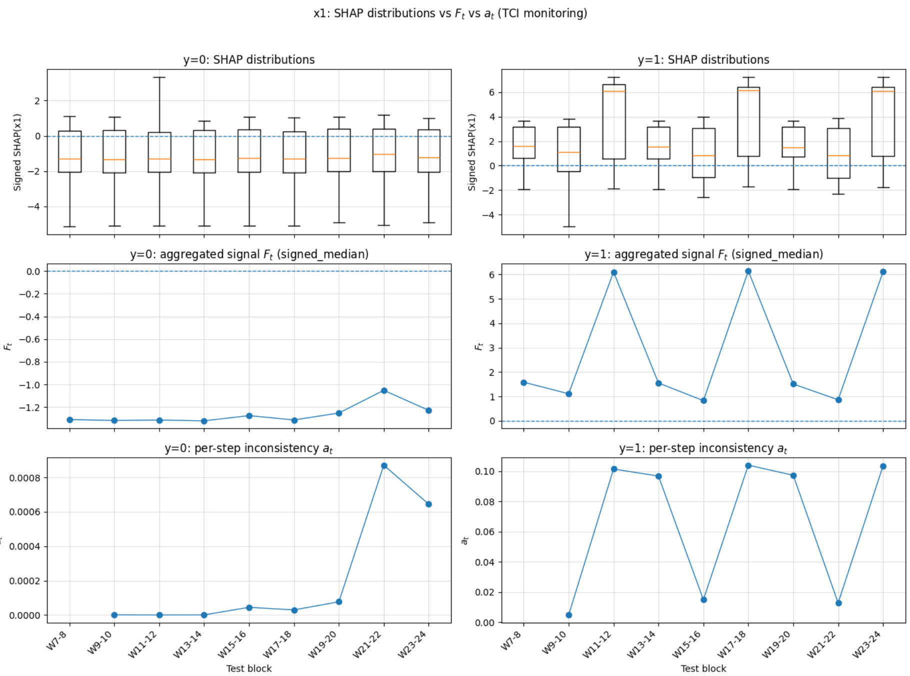

Fig. 3: Boxplots and medians of aggregated SHAP attributions for feature x1 over time steps, and per-step inconsistency, regarding stable (left plots) and unstable (right plots) classes generated using the XGBoost model.

A set of candidate summaries was generated using the linguistic quantifiers "around half" and "majority". Then, a subset of representative summaries was selected based on their degree of support and temporal relevance. For each selected summary, TCI, the mean degree of truth, and the mean degree of support are reported in Table 2. As observed, the characteristics of the XGBoost model and neural network-based summaries are similar. The most noticeable difference between the TCI for the two can be observed for the $S16$ summary, with a value of 0.58 for XGBoost and 0.46 for the neural network.

Fig. 5 and Fig. 6 illustrate the explanatory signals and per-step inconsistency contributions for the three summaries with the highest TCI generated using the XGBoost model. A comparison of SHAP-based signals and FLS-based explanatory streams reveals that the latter exhibit more intricate temporal behavior, indicative of interactions between multiple features. While the recurring structure of the simulated data remains visible, the explanatory signals no longer align directly with single-feature distributional changes.

It is worth noting that among the selected summaries, summary $S3$ most closely follows the expected temporal structure induced by the data-generation process. In contrast, the remaining summaries display more irregular behavior.

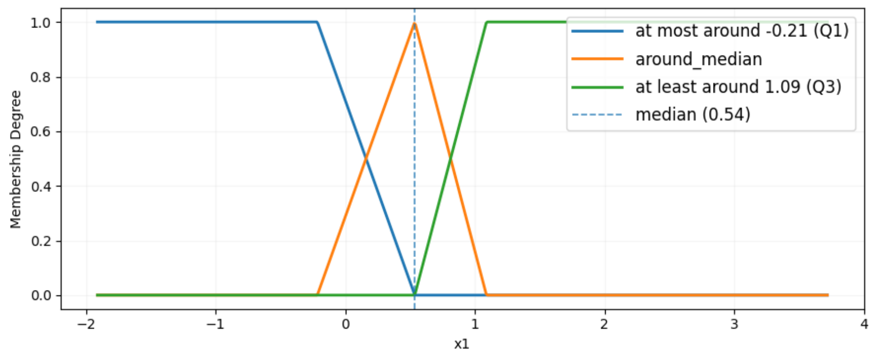

Fig. 4: Fuzzy membership functions for quartile-based linguistic expressions for feature x_1.

Table 2: Fuzzy linguistic summaries with quality measures. Quantifiers are color-coded (“majority": blue, “around half": orange); contribution polarity is indicated by color (“positive": green, “negative": red). Contributions were calculated for two alternative models (XGBoost and NN) for comparative purposes. TCI: temporal consistency index, DoT: degree of truth, DoS: degree of support.

id	**Linguistic summary**	**Model**	**TCI**	**DoT**	**DoS**
S2	For majority of obs. with x_1 at least around Q_3 and x_2 at least around Q_3, x_1 has a positive contribution to unstable class prediction.	XGB	0.44	0.33	0.15
		NN	0.44	0.33	0.15
S1	For around half of obs. with x_1 at least around Q_3 and x_3 at least around Q_3, x_3 has a positive contribution to unstable class prediction.	XGB	0.52	0.64	0.22
		NN	0.47	0.53	0.22
S16	For majority of obs. with x_1 at least around Q_3 and x_3 at least around Q_3, x_3 has a positive contribution to unstable class prediction.	XGB	0.58	0.24	0.22
		NN	0.46	0.36	0.22
S18	For around half of obs. with x_1 at most around Q_1 and x_2 around median, x_1 has a negative contribution to unstable class prediction.	XGB	0.95	0.82	0.08
		NN	0.98	1.00	0.10
S14	For around half of obs. with x_1 at most around Q_1 and x_2 around median, x_1 has a negative contribution to unstable class prediction.	XGB	0.95	0.69	0.10
		NN	0.98	1.00	0.10
S3	For around half of obs. with x_1 at most around Q_1 and x_2 at least around Q_3, x_2 has a negative contribution to unstable class prediction.	XGB	0.97	1.00	0.12
		NN	0.96	1.00	0.10

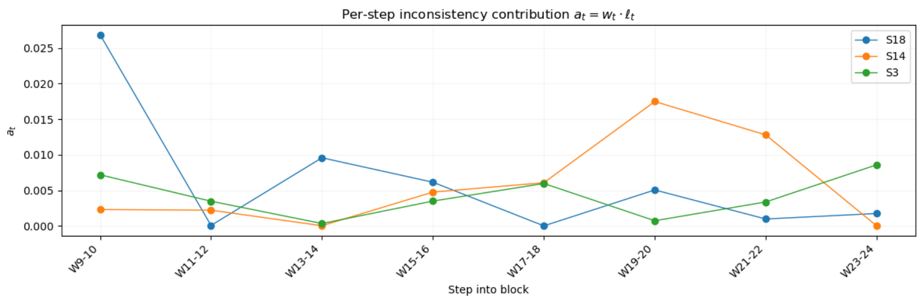

Fig. 5: Explanatory signal calculated as the product of the Degree of Truth (DoT) and Degree of Support (DoS) over time for three selected fuzzy linguistic summaries generated using the XGBoost model.

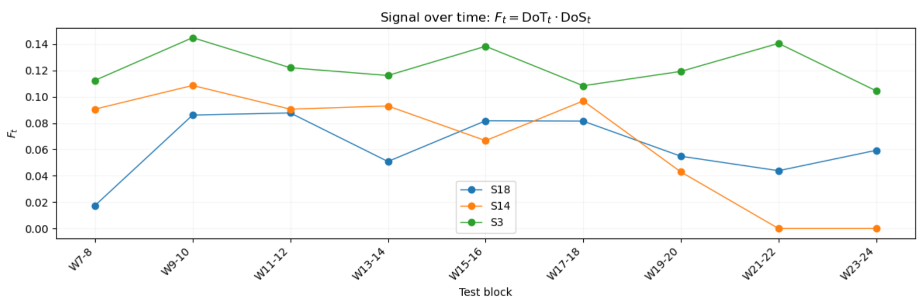

Fig. 6: Per-step inconsistency calculated on explanatory signal calculated as the product of the Degree of Truth (DoT) and Degree of Support (DoS) for three selected fuzzy linguistic summaries generated using the XGBoost model.

Nevertheless, the proposed framework remains capable of identifying periods of elevated inconsistency, thereby demonstrating its applicability to higher-level, relational explanations. In this illustrative example, the specific choice of the predictive model is not central. Any standard supervised classifier can be used, as the proposed consistency assessment operates on explanation streams independently of the underlying model.

Whilst we have focused on feature-based explanations obtained via SHAP in this paper, it should be noted that the proposed framework is equally applicable to other explanation paradigms. In particular, fuzzy rule-based explanations produced by the GUAJE framework [16] yield time-indexed explanatory quantities such as rule activation or support, which can be treated as explanatory signals and assessed using the same temporal consistency index. A systematic comparison between SHAP-based, fuzzy linguistic summary-based, and GUAJE-based explanation streams as additional symbolic explanatory processes is reserved for future research.

5 Conclusions and Future Work

The primary contributions of this work are threefold. Firstly, a temporal consistency index for explanatory signals is introduced. This index aims to quantify deviations from expected temporal behavior in a model-agnostic and explanation-agnostic manner. The proposed index operates directly on explanation streams, thus negating the need for an underlying predictive model or explanation method. This enables a unified assessment of explanation reliability over time. Secondly, we have seen that temporal consistency constitutes an additional and complementary dimension for model evaluation, revealing differences between explanations that are not captured by predictive performance alone. By employing controlled temporal scenarios, we demonstrate that explanations corresponding to different features, rules, or models can exhibit markedly different consistency profiles under non-stationary data distributions. Thirdly, the practical relevance of the proposed index is illustrated using alternative fuzzy information granules, namely fuzzy linguistic summaries. It is demonstrated how explanation consistency can inform model selection and deployment decisions by exposing trade-offs between predictive accuracy, model adaptivity, and explanation reliability in evolving environments.

Future work will extend the current pipeline to additional model classes, incorporating fuzzy rule-based predictors implemented within the GUAJE framework, probabilistic models, neural networks and logistic regression.

Acknowledgement. The project "ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech" (FENG.02.02-IP.05-0302/23) is carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG). Jose M. Alonso-Moral recognizes the support of the Galician Ministry of Education, Science, Universities and Professional Training (CiTIUS 2024-2027 ED431G-2023/04 grant) and the European Union (European Regional Development Fund - ERDF), the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033/) through MAIXAI4TRUST PID2024-157680NB-I00 project, and the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101134894 (SOSfood project).

References

1. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Ser, J.D., Díaz-Rodríguez, N., Herrera, F.: Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion (2023). https://doi.org/10.1016/j.inffus.2023.101805
2. Basseville, M., Nikiforov, I.: Detection of Abrupt Change Theory and Application, vol. 15 (04 1993)

3. Breiman, L.: Statistical modeling: The two cultures. Statistical Science **16**(3), 199—-231 (2001). https://doi.org/10.1214/ss/1009213726
4. Casalino, G., Castellano, G., Kaczmarek-Majer, K., Schicchi, D., Taibi, D., Zaza, G.: Evolving fuzzy classification for human-centered explainable learning analytics in virtual environments. Evolving Systems (2025)
5. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. KDD '16, ACM (Aug 2016). https://doi.org/10.1145/2939672.2939785
6. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017), https://arxiv.org/abs/1702.08608
7. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, H.: A survey on concept drift adaptation. ACM Computing Surveys (CSUR) **46** (04 2014). https://doi.org/10.1145/2523813
8. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models (2018), https://arxiv.org/abs/1802.01933
9. Kaczmarek-Majer, K.: Explaining neural networks in natural language. Introduction to Intricate Artificial Psychology with Python edited by P. Watson, H. Farahani, T. Bezdan (2026)
10. Lipton, Z.C.: The mythos of model interpretability (2017), https://arxiv.org/abs/1606.03490
11. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: Explainable ai for trees: From local explanations to global understanding (2019), https://arxiv.org/abs/1905.04610
12. Mariotti, E., Alonso-Moral, J.M., Gatt, A.: Measuring model understandability by means of shapley additive explanations. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–8 (2022). https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882773
13. Miś, K., Kaczmarek-Majer, K., Baczyński, M.: Fuzzy linguistic summaries and the double negation property. TechRxiv (2025)
14. Molnar, C., Casalicchio, G., Bischl, B.: Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges, p. 417–431. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-65965-3_28
15. Ostrowski, M.: The TCI repository (2025), https://github.com/Zylaz/StabilityMetric
16. Pancho, D., Alonso, J., Magdalena, L.: Quest for interpretability-accuracy trade-off supported by fingrams into the fuzzy modeling tool guaje. International Journal of Computational Intelligence Systems **6**, 46–60 (06 2013). https://doi.org/10.1080/18756891.2013.818189
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay: Scikit-learn: Machine learning in python (2018), https://arxiv.org/abs/1201.0490
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier (2016), https://arxiv.org/abs/1602.04938
19. Rojat, T., Puget, R., Filliat, D., Ser, J.D., Gelin, R., Díaz-Rodríguez, N.: Explainable artificial intelligence (XAI) on timeseries data: A survey (2021), https://arxiv.org/abs/2104.00950

20. Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., Katta, S., Donnelly, J., Chen, H., Boner, Z.: Position: amazing things come from having many good models. In: Proceedings of the 41st International Conference on Machine Learning (ICML). JMLR.org (2024)
21. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017), https://arxiv.org/abs/1703.01365
22. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning **23** (11 1994). https://doi.org/10.1007/BF00116900
23. Zadeh, L.: Fuzzy sets. Information and Control **8**(3), 338–353 (1965). https://doi.org/https://doi.org/10.1016/S0019-9958(65)90241-X