Beyond Static Importance: Quantifying Stability and Distribution Drift

Marcin Ostrowski[1,*], Olgierd Hryniewicz[1]

[1]Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-147 Warsaw, Poland

Abstract

Feature importance plays a fundamental role in machine learning and serves as a cornerstone of explainable machine learning. In temporal settings, where data accumulates sequentially, the relevance of features may evolve, introducing challenges for interpretation. While temporal variation in feature importance is increasingly relevant for applications such as clinical monitoring and time-series prediction, it remains underexplored in the literature. In this paper, we propose a novel methodology for quantifying the temporal stability of local feature attributions. Our approach combines exponentially weighted moving average (EWMA) model with performance metrics. The goal is to compute a feature-wise stability metric that reflects how consistently a feature contributes to model predictions over time. To complement this, we introduce a distributional drift score based on the Wasserstein distance, capturing shifts in the underlying feature distributions. Together, these two signals form a diagnostic framework that distinguishes between shifts due to data dynamics and those arising from model behavior. We evaluate our approach on a simulated dataset reflecting mental health monitoring scenario, as well as a publicly available benchmark time-series dataset. In both cases, the proposed metrics uncover nuanced patterns of feature behavior, enabling practitioners to identify features that are not only important but also temporally reliable. Our results demonstrate that assessing both the stability of explanations and the drift of features provides a more robust foundation for trustworthy model interpretation in dynamic environments.

Keywords

Explainable AI, Feature Importance, Time Series Analysis, Shapley Values

1. Introduction

Although methods of explainable AI (XAI) have been advancing significantly in recent years [1], yet substantial challenges persist, particularly in the context of temporal data streams. While explanation methods such as SHAP or LIME offer insights into model behavior at a given point in time, they often neglect how explanations evolve as models undergo retraining or are exposed to new data. This oversight is of critical importance, particularly in the healthcare domain, where temporal consistency of model reasoning is imperative for establishing trust and ensuring usability [2, 3].

Recent efforts have begun to explore explanation dynamics in time-dependent settings [4, 5, 6], but much of the work either focuses on specific time series models or offers descriptive analyses without actionable insights. In this work, we address a particular question of how stable feature attributions are over time and what instabilities might signal.

Feature attribution stability is interpreted as the temporal consistency of a feature's importance, as measured by an explanation method. Capturing these fluctuations is expected to yield new insights into model robustness, data drift, or redundant feature use. For instance, a feature whose importance fluctuates erratically over time may signify a model that is excessively sensitive to noise or evolving distributions. While the notion of stability has been examined in static contexts, such as in feature selection [7], few methods exist for assessing and interpreting explanation stability over evolving data.

To address this gap, we propose a novel approach that quantifies fluctuations in feature importance over time, accounting for the performance and changes of predictions as more data becomes available, thereby improving the interpretability of temporal machine learning models. Moreover, this approach

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy

*Corresponding author.

Marcin.Ostrowski@ibspan.waw.pl (M. Ostrowski)

0000-0001-9877-508X (O. Hryniewicz)

takes into account the drift of features to enhance the understanding of changes in the features themselves over time. We validate our approach using a diverse set of datasets, including a case study in mental health monitoring and benchmark dataset from the UCI Machine Learning Repository [8]. Our results demonstrate that the proposed method provides valuable insights into the dynamics of feature importance over time, considering both the temporal nature of the data and model performance. The proposed metric captures the magnitude and direction of fluctuations in feature values, which, when incorporated into a broader framework, can offer a more comprehensive understanding of feature behavior and improve feature selection.

The structure of the paper is as follows. In the next section, we present a brief description of related works. This is followed by a presentation of the proposed approach in Section 3. The experimental results using simulated data and benchmark datasets are presented in Section 4. Finally, Section 5 outlines the conclusions and discusses directions for future work.

2. Related Work

Comprehension of the role that input features play in machine learning models' predictions constitutes a fundamental principle of explainable AI [9]. These approaches can be broadly categorised into two distinct classifications: model-specific and model-agnostic. The efficacy of model-specific methods is contingent upon the utilization of internal model structures for the determination of importance. To illustrate, decision trees and ensemble models, including random forests and gradient boosting machines, are known to provide feature importance based on criteria such as Gini gain or information gain [10]. In a similar vein, linear models employ a ranking system that prioritizes features based on the magnitude of their coefficients [11]. However, the capacity for cross-model comparisons or generalisations is limited.

On the contrary, model-agnostic methods are characterised by their ability to offer greater flexibility. Permutation importance is a method of assessing a feature's relevance. It does so by evaluating the impact on model performance when feature values are randomly shuffled [12]. Despite its extensive utilization, the method is susceptible to collinearity, a factor that frequently results in an underestimation of the significance of correlated features [13]. Shapley values [14], are predicated on the principles of cooperative game theory. These values function to distribute prediction contributions among features in a equitable manner. The SHAP framework [15] has been instrumental in facilitating the calculation of Shapley values for intricate models, thereby establishing itself as a prevalent instrument for both local and global explanations.

While these approaches have been the focus of extensive research for static datasets, their application to temporal or sequential data remains limited. The majority of explanation methods treat time steps independently, neglecting to consider how the importance of features evolves or fluctuates over time [16, 17]. Recent works have begun to address this issue. For instance, Rojat et al. [4] propose temporally-aware feature attribution for time series forecasting models. Arsenault et al. [5] investigated the impact of temporal windowing on explanation variance. The extant literature indicates that explanation stability is influenced by a number of factors, including data drift, modeling choices, retraining frequency, and attribution methods. However, there is a paucity of methods that attempt to quantify these dynamics comprehensively, especially in model-agnostic settings.

Concurrently, the evaluation of explanation methods has emerged as a prominent research subject in the field [18, 19]. A variety of metrics have been proposed to evaluate the quality of explanations, including stability, fidelity, consistency, and sensitivity. However, these metrics frequently operate within a static framework and cannot effectively address the temporal characteristics inherent in data or the continual refinement of models over extended periods. Moreover, as noted by recent critiques [20, 21], explanation reliability can be compromised by issues such as feature redundancy, model non-identifiability, and attribution instability. For example, in the presence of collinear features, shifts in attribution do not necessarily reflect genuine model drift but may simply reflect equivalent representations – a core challenge our framework aims to accommodate.

In this work, we quantify the evolution of feature attributions through two complementary metrics: a stability metric and a distribution drift score. Unlike static assessments, our approach decomposes explanation dynamics into interpretable signals that reflect both model behavior and data characteristics over time. The stability metric captures deviations of feature importance from a smoothed historical baseline, weighted by model performance, thereby accounting for fluctuations in predictive reliability. In parallel, the distribution drift score quantifies changes in the underlying feature distributions using the Wasserstein distance.

3. Assessing Variations in Temporal Datasets

In this section, we will present two complementary metrics developed to capture the temporal dynamics inherent in the feature importances of multivariate time-series data. These metrics include a stability metric based on Exponentially Weighted Moving Average (EWMA) and a drift score based on Wasserstein distance. Collectively, these measures facilitate comprehensive monitoring of the consistency and distributional evolution of explanatory signals over time.

Let $\mathbf{X} \in \mathbb{R}^{K \times T}$ denote a sample of multivariate time series with K features and T time points. Let also $\mathbf{x}_t := \mathbf{X}_{\cdot,t} \in \mathbb{R}^K$ be the vector of all feature observations at a time point $1 \leq t \leq T$, where $T > 1$. Feature importance scores are computed based on the realizations $\mathbf{x}_t$ of the time series. Let $F_t^{(j)} \in \mathbb{R}$ denote the importance of the feature $j \in \{1, 2 \ldots, K\}$ at the time point t, derived from a model-specific feature attribution method, with the full sequence over time represented by $F^{(j)} = \{F_1^{(j)}, F_2^{(j)}, \ldots, F_T^{(j)}\}$.

3.1. Stability Metric

We propose a methodology for deriving a feature-wise stability score, a metric that quantifies the discrepancy between a feature's perceived importance and its exponentially weighted historical trend.

The stability metric is then defined as:

$$SM(j) = 1 - \frac{1}{T-1} \sum_{t=2}^{T} g(w_t) \frac{\left(F_t^{(j)} - \hat{F}_t^{(j)}\right)^2}{|F_t^{(j)}| + |\hat{F}_t^{(j)}| + \epsilon}, \tag{1}$$

where:

- $g(w_t) \geq 0$ is a weight function assigned at time point $t = 2, \ldots, T$, incorporating model performance w_t of the model. If $g(w_t) = 1$, $\forall t$, no weighting is applied.
- $\widehat{F}_t^{(j)}$ is the EWMA of past feature importance scores, recursively defined as:

$$\hat{F}_t^{(j)} = \lambda \hat{F}_{t-1}^{(j)} + (1 - \lambda) F_t^{(j)}, \tag{2}$$

  where $\lambda \in [0, 1]$ is the smoothing factor controlling the influence of past values. The process is initialized as $\hat{F}_1^{(j)} = F_1^{(j)}$.
- $\epsilon > 0$.

The proposed approach ensures that importance deviations are scaled in relation to their respective magnitudes, thus preventing excessive penalization of low-importance features. A high $SM(j)$ indicates that the importance of feature j exhibits a close correspondence with its historical trend, thereby suggesting temporal reliability.

The parameter λ is a component in the analysis, which governs the rate at which past observations decay. When λ is set to a larger value, it places more emphasis on the most recent data points, yielding a smoother baseline. Conversely, when λ is set to a smaller value, it gives greater weight to older observations.

The well-known EWMA is a time series analysis technique that assigns exponentially decreasing weights to past observations. This property renders the EWMA particularly useful for detecting trends and anomalies in noisy data. In contrast to the simple moving average, which assigns equal weight to all past observations, the EWMA assigns greater weight to more recent data points. This characteristic renders the EWMA more responsive to changes in the time series.

The squared difference between the feature importance value and the EWMA of past feature importance scores is aiming to capture the variability in feature importance over time. This formulation gives more weight to recent time points and models with higher predictive performance, making it particularly responsive to evolving patterns. Additionally, the weight function adjusts for model performance, assigning greater importance to better-performing models while diminishing the influence of models with lower accuracy.

The stability metric is generalizable and can be applied with any feature importance method as defined in Eq. (1). In our experiments, we selected Shapley values, which provide an attribution-based measure of feature importance.

3.2. Measuring the Distributional Drift

While the stability metric is capable of capturing fluctuations in importance, it does not have the capacity to detect whether the feature itself is undergoing a distributional shift. To address this limitation, we propose a drift score, computed as the Wasserstein-1 distance between the empirical distribution of a feature at the current time point and a reference distribution estimated from a kernel density estimate over a rolling window of past values, capturing recent trends in the feature's distribution. Both distributions are approximated using kernel density estimation.

Let $P_t^{(j)}$ denote the empirical distribution of the feature j at a time point t, and let $\widehat{P}_t^{(j)}$ denote a smoothed estimate of its past distribution. Then the drift score is defined as:

$$\mathrm{Drift}(j) = \frac{1}{T-1} \sum_{t=2}^{T} W_1(P_t^{(j)}, \widehat{P}_t^{(j)}). \tag{3}$$

In this context, $W_1(\cdot, \cdot)$ denotes the first-order Wasserstein distance originally introduced in [22] in the context of optimal transport problems. In practice, the estimation of P_t is achieved through the utilization of kernel density estimation over a designated sliding window or by computing a smoothed histogram over the designated time period. This signal captures the evolution of the underlying values of a feature, independent of its importance to the model.

Employing the stability metric $SM(j)$ and the drift score $\mathrm{Drift}(j)$, can facilitate the origin of temporal variations. For example, when high $SM(j)$ and low $\mathrm{Drift}(j)$ are observed, the feature exhibits both stable importance and a stationary distribution over time. This suggests that the feature is robust and consistently relevant for the model across time, making it a strong candidate for long-term interpretability and reliable decision-making. On the contrary, when low $SM(j)$ and high $\mathrm{Drift}(j)$ are reported, the feature's importance fluctuates over time in conjunction with significant distributional changes. This indicates that the model's reliance on the feature is adapting to underlying shifts in the data-generating process, potentially reflecting concept drift or context-sensitive importance. Another scenario would be to observe low $SM(j)$ and low $\mathrm{Drift}(j)$. The feature's distribution remains largely stable, yet its attributed importance varies. This scenario may signal model instability, such as overfitting or sensitivity to transient patterns, since fluctuations in explanatory power are not explained by input variation. Alternatively, high $SM(j)$ and high $\mathrm{Drift}(j)$ would mean that despite the feature undergoing distributional changes, its importance remains stable. This indicates model robustness to input drift, suggesting that the feature maintains its predictive role under varying conditions — a desirable trait in non-stationary environments.

4. Numerical Results

We begin by validating the proposed stability and drift metrics using a simulated dataset derived from real-world clinical observations, followed by experiments on a publicly available benchmark dataset [8]. In both cases, the input consists of multivariate time series with associated labels, where each time point corresponds to a set of feature observations and an outcome. The goal of the experiments is to assess how feature importance evolves over time, and to evaluate whether the proposed metrics — stability and distributional drift — can meaningfully capture changes in model attribution dynamics under temporal shifts in data. To this end, we apply a rolling-window training procedure using XGBoost classifiers and compute both feature importance and drift signals across time. Full implementation details and data are available in the publicly accessible repository [23].

4.1. Simulation study

Our evaluation is motivated by a real-world clinical problem concerning the remote sensor-based monitoring of bipolar disorder patients. In this particular applied problem, considering the feature's importance for a specific time point does not provide comprehensive insights into the temporal influence of the feature. If a feature's importance may change over time significantly, ranging from the most important to least important, without a clear pattern, the inference based on such a feature might not be as reliable as assumed. Our approach aims to identify features that demonstrate temporal stability in both importance and distributional behavior, thus enhancing interpretability and trust in model-based inference.

The real-world dataset comprises acoustic and psychiatric data collected from a patient diagnosed with bipolar disorder. For further details see [24] with the protocol of this clinical study. Acoustic features describe the manner of speaking with physical descriptors such as shimmer, jitter, and energy extracted from patients' speech. Additionally, each feature vector is associated with the patient's mental state at the time of the recording. Bipolar disorder is a serious mental illness characterized by fluctuations from depressive through euthymic to manic states. Previous research confirms that acoustic features extracted from speech serve as valid markers for assessing the severity of manic and depressive symptoms[25]. In this work, we aim to assess the stability of the feature's importance in time.

To reduce complexity and enable stable classification under limited data, the original multiclass labeling was binarized mapping states into two categories: euthymia, which is considered the healthy state, and non-euthymia.

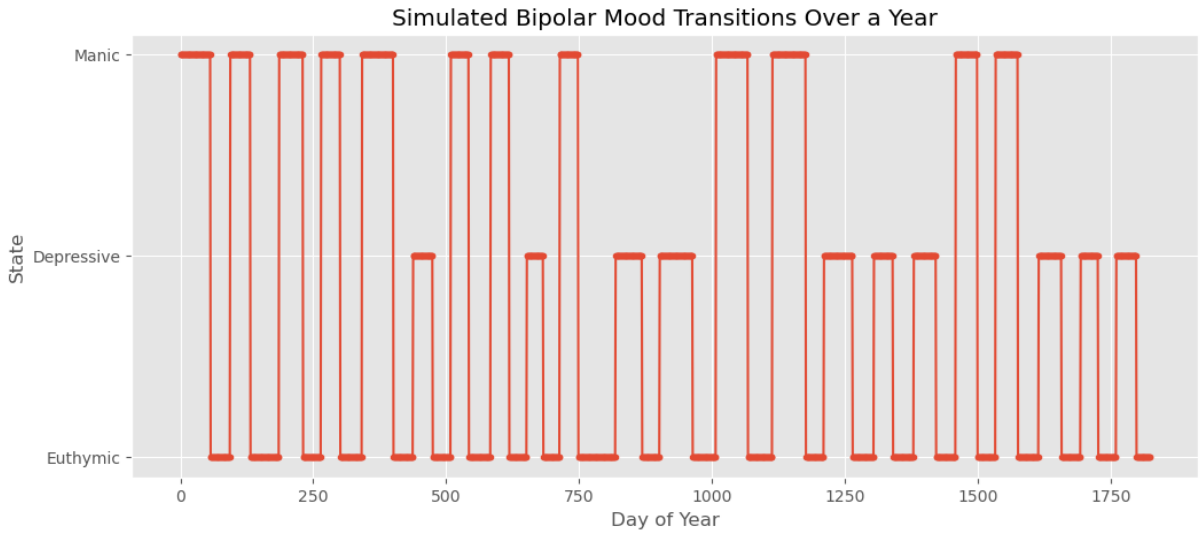


Figure 1: Changes in mental state over time for the simulated dataset.

Considering the limited size of the real data and its complex nature, we simulated a larger, controlled dataset that permits more rigorous and repeatable validation of the proposed metrics. The resulting dataset is publicly available for further investigation [23]. Fig. 1 shows the simulated changes in mental state over time. We selected four voice characteristics – jitter, shimmer, energy, and voice pitch – and used their means and standard deviations from the original dataset to model two distributions: a base distribution representing prior knowledge of mental state classification and a patient-specific distribution incorporating slight variations to reflect individual voice characteristics. Fig. 2 illustrates distributions for two exemplary simulated variables. More figures and details can be found in the publicly available repository [23].

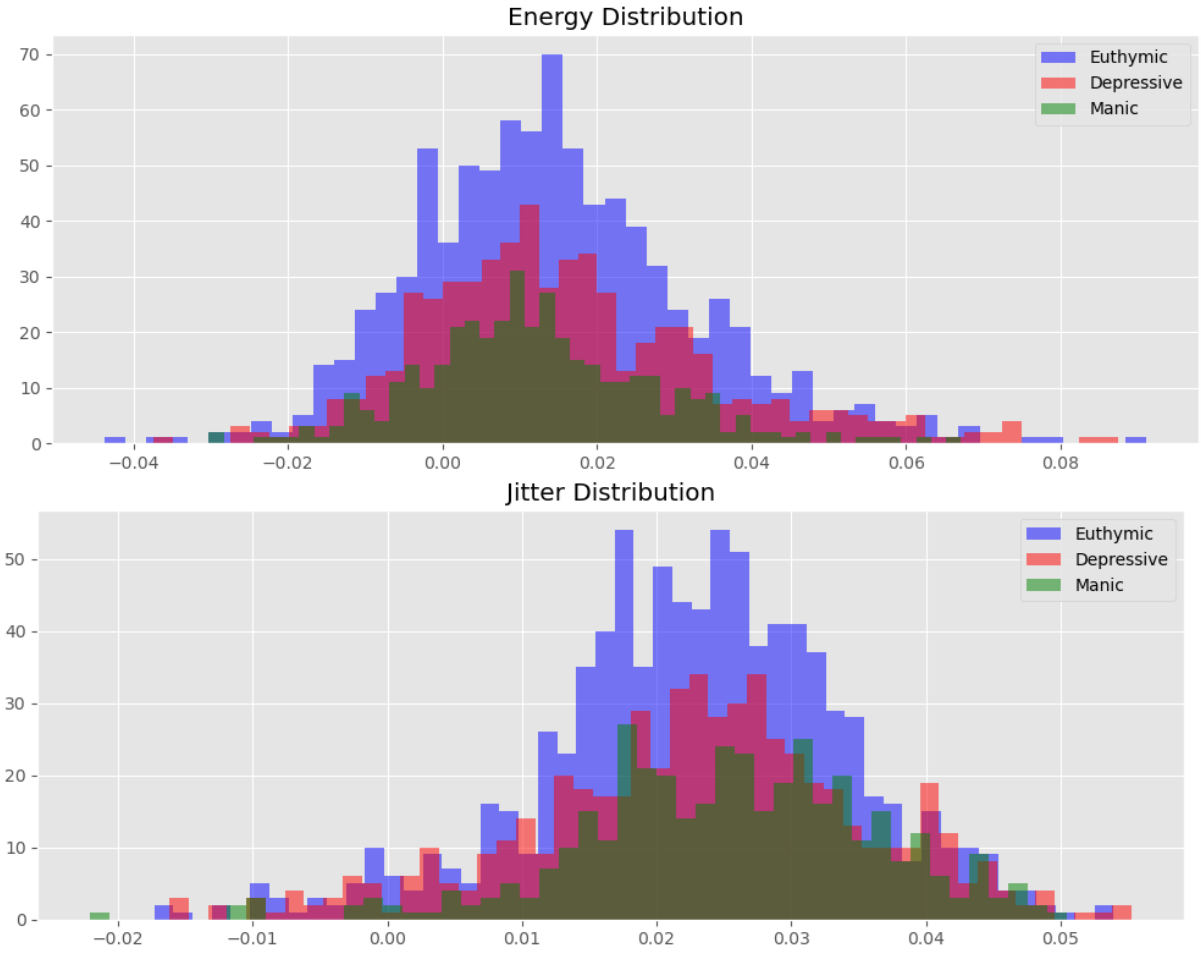


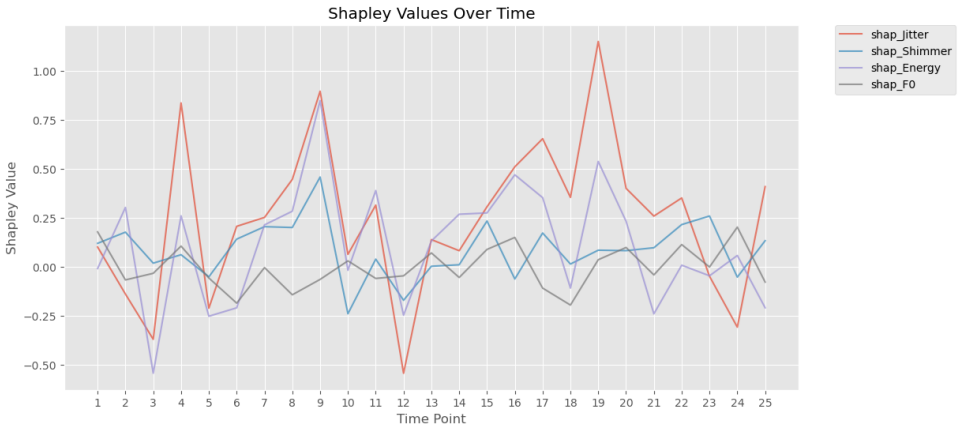
Figure 2: Distributions of the voice characteristics (energy and jitter) for data simulated for five years.

We simulated observations from patient-specific distribution over five-year period. Each observation corresponds to a specific day, and the dataset was divided into two categories of periods: ground truth periods, which spanned nine days around a psychiatrist appointment where the patient's mental state was known, and inter-appointment periods, representing the intervals between psychiatrist visits where the patient's mental state was inferred. Each inter-appointment period consisted of 64 days, resulting in five meetings per year.

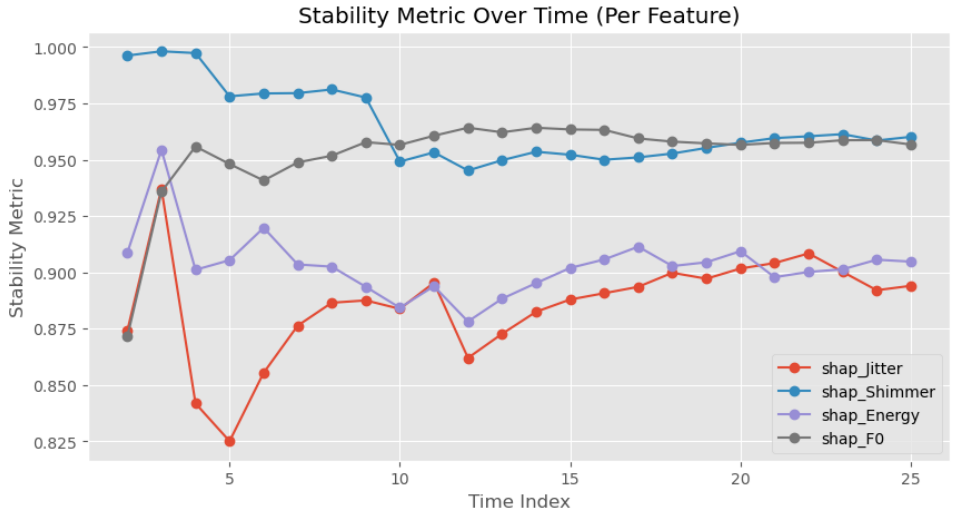
Then, we simulated 180 observations from the base distribution, with 60 observations from each state. This dataset was used as the prior knowledge dataset, while the five-year dataset included 1 825 observations with varying label distributions.

As the next step, we trained XGBoost classifiers [26] from the XGBoost package in a sequential manner. The first model was trained on the first ground truth period (9 observations) and tested on the subsequent inter-appointment period (64 observations). The second model was trained on the first two ground truth periods combined with the first inter-appointment period (a total of 82 observations),

while the test set consisted of the second inter-appointment period (64 observations). Next models followed the same pattern, each incorporating more prior data.



(a)



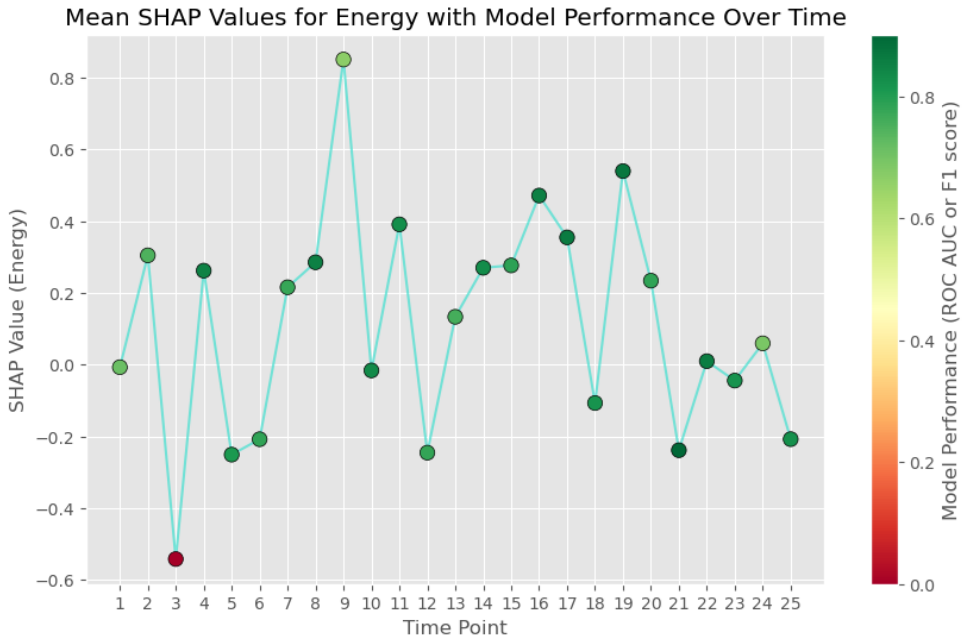
(b)

Figure 3: (a) Changes over time in mean Shapley values for voice features in simulated dataset over five years and (b) stability metric over time for the simulated dataset using Shapley values as feature importance metric in Eq. (1) and $\lambda = 0.7$ in Eq. (2).

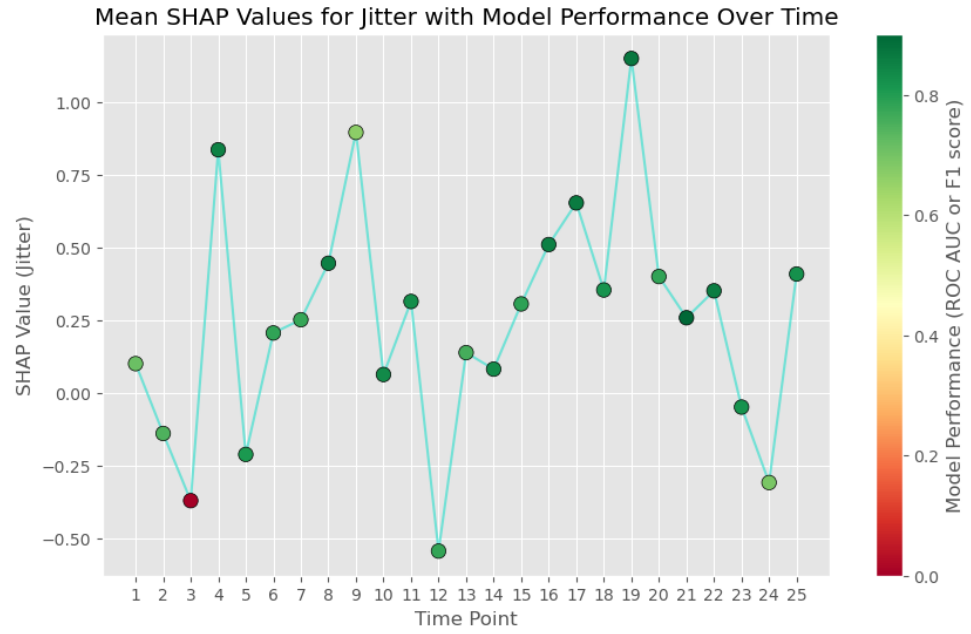
For each window, model performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC), or the $F1$ score if ROC AUC was unavailable, from the Scikit-learn package. Distribution drift was calculated for each window using Kernel Density Estimation. The drift between the current window and a smoothed estimate of the past distribution was then computed using the Wasserstein distance. Shapley values, computed using the SHAP package's Explainer interface, were calculated for the prediction of the class representing the non-euthymic state. To calculate the stability metric, a vaue of $\lambda = 0.7$ was used in all experiments (see Eq. (2)).

For the simulated medical dataset, Fig. 3 presents Shapley values over time for five-year simulations, along with stability metrics incorporating model weighting by performance. Most stability metric values exceed 0.875, suggesting either minimal fluctuations in Shapley values or that the corresponding features have little influence on the model's predictions. However, a closer inspection reveals that jitter, despite having the highest absolute Shapley values, exhibits significant variability in both magnitude and direction over time. This instability is captured in the stability metric, where jitter consistently shows the lowest stability values across both simulation periods.

Fig. 4 illustrates temporal changes in Shapley values for jitter and energy, alongside model per-



(a)



(b)

Figure 4: (a) Changes over time in mean Shapley values for voice features in simulated dataset over five years and (b) stability metric over time for the simulated dataset using Shapley values as feature importance metric in Eq. (1) and $\lambda = 0.7$ in Eq. (2).

formance at each time point. Notably, energy shows considerable fluctuations in its Shapley values, but these gradually stabilize at low positive values, with occasional negative spikes. This leads to an approximately 0.9 stability metric. On the other hand, jitter experiences both magnitude and direction fluctuations before stabilizing at positive values. However, due to more significant fluctuations and a later stabilization, jitter achieves a lower stability metric of around 0.85, before eventually reaching 0.9. The performance-based weighting smooths some Shapley value changes, particularly between time points 2, 3, and 4 where energy's Shapley values show noticeable changes. Notably, feature F0 maintains a high stability metric, reflecting its consistently low Shapley values relative to the other

features throughout the time period.

Table 1 presents the stability metric and distribution drift values for the features considered in the simulated dataset. Since no universal thresholds exist for interpreting these metrics, we analyze them relative to one another. The cell color intensity reflects these internal comparisons and helps surface outliers and patterns.

Table 1
Stability metrics and distribution drift for each of the features in the simulated dataset. Cell color intensity reflects the relative magnitude of values within each column.

Feature	Stability Metric	Distribution Drift
Jitter	0.89	0.03
Shimmer	0.96	0.03
Energy	0.90	0.02
F0	0.96	0.26

We observe that jitter, shimmer and energy exhibit relatively low distribution drift (all below 0.03) and moderate-to-high stability scores, ranging from 0.89 to 0.96. These features maintain consistent importance over time and stem from relatively stable data distributions – indicating strong potential as interpretable and dependable predictors in a temporal context.

In contrast, the F0 feature stands out with the highest distribution drift (0.26), substantially exceeding that of the other features. Despite this, it shares the highest stability score (0.96), indicating that although the underlying distribution of this feature changes substantially, its relevance to model predictions remains stable – suggesting model robustness or invariance to feature drift.

A particularly instructive case is jitter. Although it frequently appears as one of the most influential features (based on raw Shapley values), its lower stability metric indicates substantial temporal variability. This highlights the risk of over-interpreting raw importance scores without considering temporal consistency. Features like energy and shimmer, with slightly lower but more stable contributions, may offer more reliable insight for longitudinal interpretation or downstream decision-making.

This example illustrates the core utility of the proposed approach: it enables a nuanced decomposition of explanation quality, capturing both temporal stability and drift sensitivity. Such decompositions are essential when local explanations are used to guide clinical insight or intervention strategy. Traditional feature importance analysis cannot offer this level of granularity – particularly when dealing with sequential or drifting data.

A current limitation of the analysis is the omission of long-term average Shapley trends, which could offer additional insights into persistent feature relevance. This omission could result in incomplete inferences about the features, as the evolution of the overall importance of features across the entire time period is not taken into account. The incorporation of this information could facilitate a more comprehensive understanding of feature behavior and stability.

4.2. Experiments for Benchmark Dataset

To assess the generalizability of our proposed metrics beyond the medical domain, we extend our experimental framework to a publicly available time-series dataset, namely the Rocket League dataset[8]. This dataset consisted of $7\,189$ observations with 16 explanatory features and was divided into 44 contiguous gameplay segments, each corresponding to a distinct phase within a Rocket League match. The dataset was designed for binary classification.

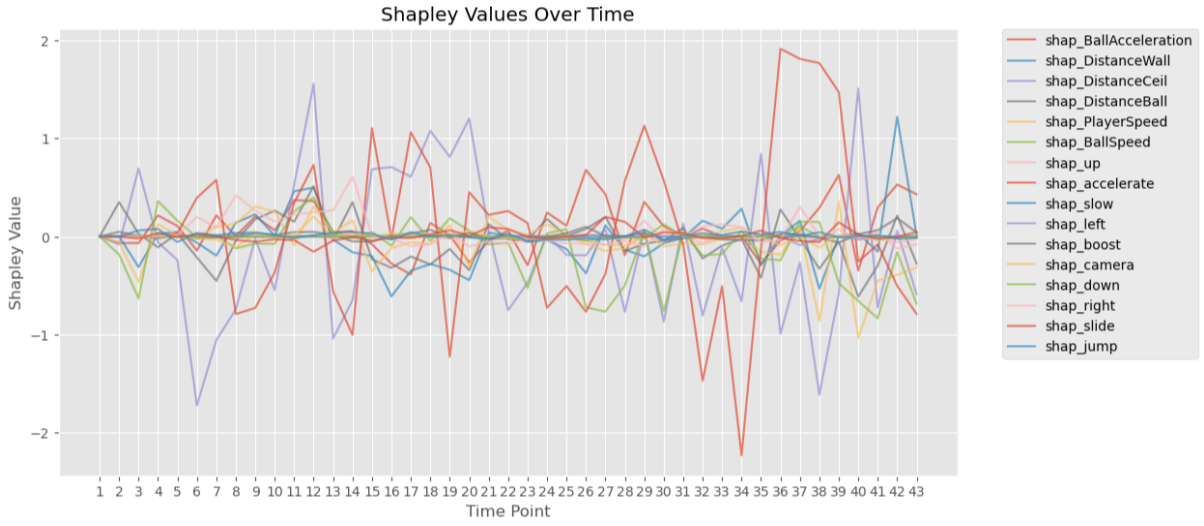
The temporal division strategy and the structural framework (XGBoost) utilized in the medical simulation were employed in this investigation. The performance-weighted stability metric and distribution drift computations were also adopted.

Fig. 5 illustrates temporal changes in mean Shapley values and the corresponding stability metrics for selected features in the Rocket League dataset. We highlight three features that reflect diverse dynamic behaviors over time.

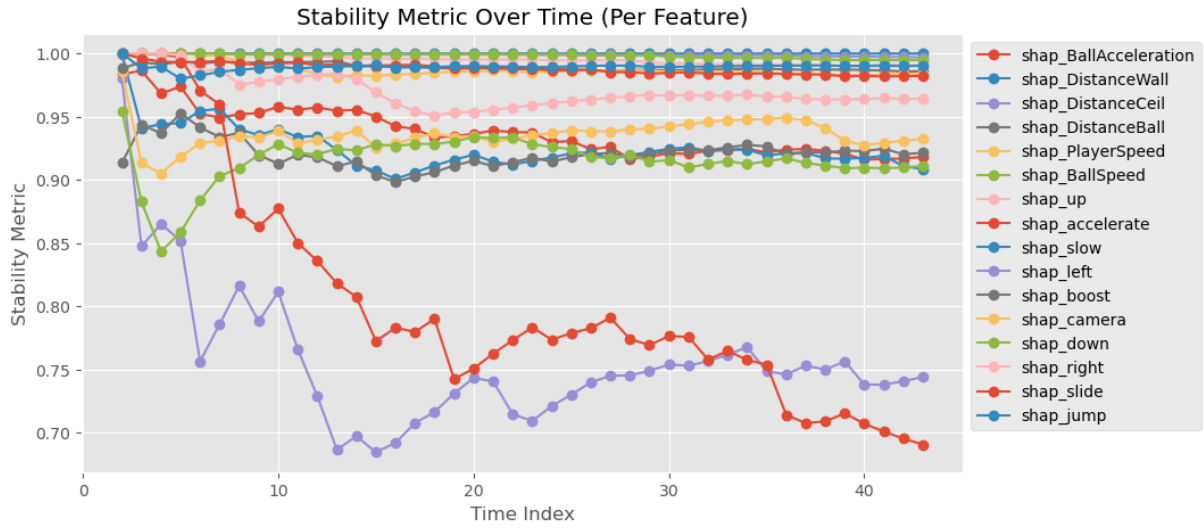
First, consider `BallAcceleration`. While its Shapley values oscillate in both magnitude and sign during the early intervals, the feature eventually stabilizes with increasingly consistent positive contributions. Despite initial fluctuations, the combination of late-stage consistency and strong positive attribution boosts its overall stability metric to around 0.92, placing it in the relative high-stability category.

In contrast, `accelerate` shows highly variable Shapley values with no consistent trend – swinging from strong positive to negative contributions across different intervals. This behavior results in a low stability metric of approximately 0.69, even though its distribution drift is minimal. Such a combination – low stability with low drift – may suggest overfitting or excessive context dependence, limiting the interpretability and reliability of this feature.

Finally, `PlayerSpeed` provides a third example, showing persistent relevance across the entire timeline. While its SHAP values exhibit some fluctuation in magnitude, their sign and overall importance remain stable. This yields a relatively high stability score (above 0.93) despite the feature undergoing substantial distributional shift, as reflected in one of the highest drift values in the dataset. This resilience implies that the model reliably incorporates `PlayerSpeed` despite changes in its input distribution – a sign of potential robustness or invariance.



(a)



(b)

Figure 5: (a) Changes over time in mean Shapley values for features in Rocket League dataset and (b) values of stability metric for Rocket League dataset, with weighting by performance in Eq. (1) and $\lambda = 0.7$ in Eq. (2).

Table 2 summarizes the stability and distribution drift scores. As in the previous case, the absence of universal thresholds necessitates relative interpretation within this dataset. The color coding provides a visual anchor for quickly identifying anomalies or outliers.

Compared to the simulated dataset, drift scores in this domain span a much broader range – from near-zero (e.g., down, `left`) to extreme values exceeding 38,000 (e.g., `BallSpeed`).

An especially informative case is `accelerate`. Despite occasional influence across time steps, its low stability and minimal drift make it a poor candidate for interpretation – highlighting the risk of overemphasizing volatile features in temporal contexts. In contrast, `PlayerSpeed` and `BallSpeed` demonstrate that strong, stable attributions can persist even amid significant input drift, suggesting that the model can reliably leverage these features under changing conditions. Finally, features such as `up`, `slow`, and `left`, which exhibit both high stability and low drift, stand out as particularly robust explanatory signals. Together, these findings underscore the key contribution of the proposed framework: by integrating attribution stability with distributional drift, it enables a deeper, time-sensitive assessment of explanation quality – beyond what traditional feature importance scores can offer.

Table 2
Stability metrics and distribution drift for each of the features in the Rocket League dataset. Cell colors indicate relative magnitude within each column.

Feature	Stability Metric	Distribution Drift
BallAcceleration	0.92	14204.22
DistanceWall	0.91	11657.49
DistanceCeil	0.74	432.07
DistanceBall	0.92	850.12
PlayerSpeed	0.93	33042.16
BallSpeed	0.91	38073.88
up	0.99	0.05
accelerate	0.69	0.17
slow	0.99	0.05
left	0.99	0.05
boost	0.99	0.14
camera	0.98	0.09
down	0.99	0.04
right	0.96	0.11
slide	0.98	0.18
jump	0.99	0.16

5. Conclusion and Future Work

This study introduces a novel framework for evaluating the temporal stability of feature importance, while jointly accounting for distributional drift in the underlying data. We operationalize this idea through a performance-weighted stability metric and a drift score based on the Wasserstein distance, applied to local explanations derived from Shapley values. By incorporating exponentially weighted moving averages and performance-based weights, the proposed stability metric captures deviations in feature importance over time in a calibrated manner. The drift score complements this by quantifying changes in the empirical distribution of each feature, thereby disentangling model instability from data drift.

Our empirical validation on both a simulated clinical dataset and a publicly available benchmark time series demonstrates the utility of the framework in diagnosing and interpreting model behavior over time. The results affirm that reliable model interpretation cannot rely solely on feature importance magnitudes; instead, it should incorporate both temporal consistency and feature dynamics.

While the proposed metrics offer valuable diagnostic insights, their full utility emerges when integrated with complementary analytical signals. For instance, temporal autocorrelation or variance decomposition could provide further context about the persistence and volatility of individual features. Additionally, a formal framework for aggregating multiple explanation signals – stability, drift, rank variance, and uncertainty – could further improve understanding of the modeling process under temporal shifts and its outputs. Future work will extend the analysis to additional feature importance metrics, further exploring their impact on stability evaluation. Finally, future research will focus on refining the weighting schemes in the stability metric, extending the approach to additional feature attribution methods beyond Shapley values, evaluating model robustness under adversarial or synthetic distribution shifts, and applying the framework to other domains. We also plan to release an open-source toolkit implementing these metrics to facilitate adoption and further experimentation.

Acknowledgments

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).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, F. Giannotti, A survey of methods for explaining black box models (2018). doi:10.48550/ARXIV.1802.01933. arXiv:1802.01933.

[2] P. Gohel, P. Singh, M. Mohanty, Explainable ai: current status and future directions (2021). doi:10.48550/ARXIV.2107.07045. arXiv:2107.07045.

[3] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning – a brief history, state-of-the-art and challenges, Koprinska I. et al. (eds) ECML PKDD 2020 Workshops. ECML PKDD 2020. Communications in Computer and Information Science, vol 1323. Springer, Cham (2020) 417–431. doi:10.1007/978-3-030-65965-3_28. arXiv:2010.09337.

[4] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, N. Díaz-Rodríguez, Explainable artificial intelligence (xai) on timeseries data: A survey (2021). doi:10.48550/ARXIV.2104.00950. arXiv:2104.00950.

[5] P.-D. Arsenault, S. Wang, J.-M. Patenande, A survey of explainable artificial intelligence (xai) in financial time series forecasting (2024). doi:10.48550/ARXIV.2407.15909. arXiv:2407.15909.

[6] T. T. Nguyen, T. Le Nguyen, G. Ifrim, Robust explainer recommendation for time series classification, Data Mining and Knowledge Discovery 38 (2024) 3372–3413. URL: http://dx.doi.org/10.1007/s10618-024-01045-8. doi:10.1007/s10618-024-01045-8.

[7] M. C. Barbieri, B. I. Grisci, M. Dorn, Analysis and comparison of feature selection methods towards performance and stability, Expert Systems with Applications 249 (2024) 123667. doi:10.1016/j.eswa.2024.123667.

[8] R. Mathonat, Rocket League Skillshots, UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C5S035.

[9] Z. C. Lipton, The mythos of model interpretability (2016). doi:10.48550/ARXIV.1606.03490. arXiv:1606.03490.

[10] J. H. Friedman, Greedy function approximation: A gradient boosting machine., The Annals of Statistics 29 (2001). doi:10.1214/aos/1013203451.

[11] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer New York, 2009. doi:10.1007/978-0-387-84858-7.

[12] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, Journal of Machine Learning Research 20 (177), 1-81, 2019 (2018). doi:10.48550/ARXIV.1801.01489. arXiv:1801.01489.

[13] C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC Bioinformatics 8 (2007). doi:10.1186/1471-2105-8-25.

[14] L. S. Shapley, 17. A Value for n-Person Games, Princeton University Press, 1953, pp. 307–318. doi:10.1515/9781400881970-018.

[15] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions (2017). doi:10.48550/ARXIV.1705.07874. arXiv:1705.07874.

[16] M. Villani, J. Lockhart, D. Magazzeni, Feature importance for time series data: Improving kernelshap (2022). doi:10.48550/ARXIV.2210.02176. arXiv:2210.02176.

[17] K. K. Leung, C. Rooke, J. Smith, S. Zuberi, M. Volkovs, Temporal dependencies in feature importance for time series predictions (2021). doi:10.48550/ARXIV.2107.14317. arXiv:2107.14317.

[18] M. Pawlicki, A. Pawlicka, F. Uccello, S. Szelest, S. D'Antonio, R. Kozik, M. Choraś, Evaluating the necessity of the multiple metrics for assessing explainable ai: A critical examination, Neurocomputing 602 (2024) 128282. doi:10.1016/j.neucom.2024.128282.

[19] E. Mariotti, J. M. Alonso-Moral, A. Gatt, Measuring model understandability by means of shapley additive explanations, in: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2022, pp. 1–8. doi:10.1109/fuzz-ieee55066.2022.9882773.

[20] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods (2018). doi:10.48550/ARXIV.1806.08049. arXiv:1806.08049.

[21] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20, ACM, 2020, pp. 180–186. doi:10.1145/3375627.3375830.

[22] L. V. Kantorovich, Mathematical methods of organizing and planning production, Management Science 6 (1960) 366–422. doi:10.1287/mnsc.6.4.366.

[23] M. Ostrowski, Stability metric github repository, 2025. URL: https://github.com/Zylaz/StabilityMetric.

[24] M. Sokół-Szawłowska, O. Kamińska, M. Sochacka, Moodmon: novel optimization of bipolar disorder monitoring through patient-driven voice parameter submission and ai technology, Advances in Psychiatry and Neurology/Postępy Psychiatrii i Neurologii 33 (2024) 230–240. URL: http://dx.doi.org/10.5114/ppn.2024.147100. doi:10.5114/ppn.2024.147100.

[25] K. Kaczmarek-Majer, M. Dominiak, A. Z. Antosik, O. Hryniewicz, O. Kamińska, K. Opara, J. Owsiński, W. Radziszewska, M. Sochacka, L. Święcicki, Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: A prospective study, Acta Psychiatrica Scandinavica 151 (2024) 358–374. doi:10.1111/acps.13735.

[26] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system (2016) 785–794. doi:10.1145/2939672.2939785. arXiv:1603.02754.