Monitoring Explanation Drift in Longitudinal NLP Pipelines: A Cross-Domain Study under Temporal Distribution Shift

Pedro Luís Carneiro Jovino[1], Ewaldo Eder Carvalho Santana[1]

Department of Computer Engineering, Maranhão State University, Brazil
LAPS (Laboratório de Análise e Processamento de Sinais)
jovinopedroluis@gmail.com, ewaldoeder@gmail.com

Abstract. This paper studies *explanation drift*: the temporal instability of ranked human-readable terms produced by fixed explanation mechanisms in longitudinal NLP pipelines. We compare corpus-level TF–IDF explanations with model-level explanations derived from logistic-regression coefficients in arXiv and PubMed abstracts from 2014 to 2024. The vocabulary is learned only from the 2014–2016 reference window and then frozen, approximating a deployed monitoring scenario. arXiv is used as the primary supervised setting, while PubMed provides a cross-domain robustness check with query-defined weak labels. Results show that model-level explanations drift more than corpus-level TF–IDF summaries. On arXiv, mean DriftScore is 0.50 for model-level explanations and 0.15 for TF–IDF, a 3.33-fold increase with strong statistical support ($p_{adj} = 0.0040$). On PubMed, the same direction appears with smaller magnitude, 0.20 versus 0.13, but is not significant after Bonferroni correction ($p_{adj} = 0.0938$). The proposed audit rule produces no automatic alerts under the strict configuration because observed drift does not strictly exceed the null 95th percentile, illustrating the distinction between measurable explanation turnover and drift that is sufficiently unusual to trigger operational review.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Machine learning algorithms**; • **Information systems** → *Data mining*.

Keywords: Explainable AI, Explanation Drift, Temporal Distribution Shift, Longitudinal NLP, Text Mining, Model Monitoring

1. INTRODUCTION

Explainable AI is commonly evaluated under static assumptions: an explanation method is selected, validated, and then treated as a stable interface between a model and its users. This assumption is fragile in NLP. Text streams evolve, terminology changes, and the distribution that gives meaning to explanatory terms may shift over time. A model can therefore remain technically interpretable while the explanation artifacts exposed to analysts become less comparable across years.

This paper studies that failure mode. We define an *explanation artifact* as a ranked set of human-readable terms generated by a fixed explanation mechanism, such as TF–IDF scores or linear-model coefficients. The question is not whether language changes in general, but whether a stable explanation interface continues to expose comparable artifacts when the underlying corpus distribution changes.

We construct a controlled longitudinal study using scientific abstracts from 2014 to 2024 in two domains. The first corpus contains arXiv abstracts from `cs.AI`, `cs.CL`, and `cs.LG`, representing the primary supervised setting with stable category labels in a fast-moving computer-science domain. The second corpus contains PubMed abstracts retrieved through biomedical query-defined categories. PubMed is not intended to be a perfectly matched supervised benchmark; instead, it is used as a cross-domain robustness check with a different terminological profile and weaker labels. This design allows

us to distinguish the main arXiv result from a domain-specific artifact while keeping the strength of the supervised claims anchored in the arXiv corpus.

We compare two explanation levels. Corpus-level explanations are obtained from yearly TF–IDF rankings. Model-level explanations are obtained from logistic-regression coefficients trained to predict annual category labels. This separation allows us to test whether model-level explanation artifacts exhibit temporal instability beyond the lexical turnover already present in the underlying corpus, without assuming that such instability is attributable to a single underlying cause.

The contribution is intentionally operational rather than algorithmic. We do not introduce a new explainer; instead, we treat explanation artifacts as time-dependent system outputs that can themselves be monitored in longitudinal NLP pipelines. Using ranked-set overlap and DriftScore, we separate corpus-level TF–IDF drift from model-level coefficient drift, test whether model-level artifacts exhibit greater temporal instability than corpus-level summaries, and define a conservative audit rule based on an empirical null distribution. The rule provides an explicit and configurable boundary for human review when explanation continuity weakens.

2. RELATED WORK

This work connects explainable AI, ML monitoring, diachronic NLP, and concept drift. Prior XAI research studies interpretability, robustness, and human-centered explanation [1; 2; 3], including instability of post-hoc explanations under perturbations [4; 5]. Long-running ML systems additionally require monitoring of distribution shift and changing data assumptions [6; 7], while diachronic NLP and concept-drift research establish that language, feature relevance, and data-generating processes evolve over time [8; 9; 12; 13]. We extend this monitoring perspective to explanation artifacts themselves, asking whether ranked explanatory terms remain comparable across time under a fixed interface.

3. DATA AND METHOD

3.1 Corpora and Labels

Table I summarizes the datasets. arXiv is the primary supervised corpus and contains abstracts from `cs.AI`, `cs.CL`, and `cs.LG`, using the original category field as the label. PubMed serves as a cross-domain robustness corpus and uses query-defined weak labels for machine learning (`pm.ML`), clinical decision support (`pm.CDS`), and natural language processing (`pm.NLP`). These PubMed groups are retrieval-based rather than gold-standard biomedical classes and are therefore used only to test whether the explanation-drift pattern is directionally preserved across domains.

Table I. Corpora used in the longitudinal experiments. PubMed labels are query-defined weak labels.

Corpus	Labels	Years	Documents	Vocabulary setting
arXiv	`cs.AI/cs.CL/cs.LG`	2014–2024	8,656	3,000 terms, 2014–2016
PubMed	`pm.ML/pm.CDS/pm.NLP`	2014–2024	9,077	3,000 terms, 2014–2016

3.2 Preprocessing and Prospective Vocabulary

Two preprocessing settings are evaluated. In the raw setting, we apply whitespace normalization and preserve surface forms. In the lemmatized setting, abstracts are lowercased, tokenized, filtered to alphabetic tokens, stripped of stopwords, and lemmatized with `en_core_web_sm`. The comparison tests whether explanation drift is only morphological or whether it persists after lexical normalization.

For each corpus and preprocessing condition, a fixed vocabulary $\mathcal{V}$ is learned from the initial temporal window $\{2014, 2015, 2016\}$ and then reused across all years. The vocabulary is learned from the

data with a maximum of 3,000 features and fixed tokenization. This prospective design is important: terms are not added adaptively in later years, and later-year terminology is not used to define the initial feature space. Therefore, observed changes in explanation artifacts reflect ranking and coefficient changes over a stable interface rather than feature-space expansion.

3.3 Explanation Artifacts

For corpus-level explanations, TF–IDF is computed separately within each year. For a document d, term w, and annual corpus D_t, the score is

$$\mathrm{TFIDF}_t(d, w) = TF_t(d, w) \log\left(\frac{|D_t|}{DF_t(w)}\right), \tag{1}$$

where $DF_t(w)$ is the number of documents in year t containing w. Scores are aggregated across documents as

$$S_t(w) = \frac{1}{|D_t|} \sum_{d \in D_t} \mathrm{TFIDF}_t(d, w). \tag{2}$$

The top-K terms ranked by $S_t(w)$ form the yearly corpus-level explanation artifact. A document-frequency ranking over the same fixed vocabulary is also computed as a lexical baseline.

For model-level explanations, a logistic-regression classifier is trained independently for each year to predict the category label. Features are TF–IDF values over the same fixed vocabulary. The model-level explanation artifact for a year is the top-K set obtained from absolute coefficient magnitudes averaged across classes. Logistic regression is used because its coefficients provide a direct and auditable explanation signal; however, coefficients can be sensitive to correlated features, which we discuss as a limitation.

3.4 Drift Score and Null Distribution

To quantify explanation instability, we compare top-K explanation sets from consecutive years. Let K_t and K_{t-1} denote the top-K explanatory terms for years t and $t-1$. We define

$$\mathrm{DriftScore}(t) = 1 - \frac{|K_t \cap K_{t-1}|}{K}. \tag{3}$$

A DriftScore of zero indicates perfect preservation of explanatory terms; a value of 0.60 means that 60% of the top-K terms changed. Unless otherwise stated, $K = 10$.

Uncertainty is estimated with 95% bootstrap confidence intervals by resampling documents within each year and recomputing top-K sets over 200 iterations. As a complementary robustness check, we compute an empirical null distribution by permuting year labels across all documents and recomputing DriftScore over 200 permutations. The null estimates what turnover would be expected from sampling variation under no real temporal structure. We use a strict audit rule: DriftScore must be greater than, not merely equal to, the 95th percentile of the null distribution.

3.5 Monitoring Rule

Algorithm 1 formalizes the operational rule. DriftScore measures explanation turnover, whereas actionability is defined relative to the empirical null: an audit is triggered only when DriftScore exceeds the selected null percentile for at least m consecutive transitions. Thus, variation compatible with the no-temporal-structure baseline is treated as expected, while persistent exceedance is considered actionable. This policy is intentionally conservative and does not assume a universal DriftScore threshold.

Algorithm 1: ExplanationDriftMonitor

Input: Temporal corpora $D_1, \ldots, D_T$; explainer E; top-K; null percentile α; minimum consecutive exceedances m

Output: Audit years A and monitoring log L

Learn vocabulary $\mathcal{V}$ from the initial reference window and freeze it
Estimate null threshold q_t for each transition by permuting year labels
$consecutive \leftarrow 0$; $A \leftarrow \emptyset$
for $t = 2$ **to** T **do**
    Extract $K_{t-1} = E(D_{t-1}, \mathcal{V})$ and $K_t = E(D_t, \mathcal{V})$
    $s_t \leftarrow 1 - |K_t \cap K_{t-1}|/K$
    if $s_t > q_t$ **then**
        $consecutive \leftarrow consecutive + 1$
    else
        $consecutive \leftarrow 0$
    if $consecutive \geq m$ **then**
        Add year t to A and trigger explanation audit
        $consecutive \leftarrow 0$
    Append $(t, s_t, q_t, consecutive)$ to L
return A, L

3.6 Statistical Testing and Temporal Transfer

To test whether model-level drift is greater than corpus-level drift, we use a paired Wilcoxon signed-rank test over yearly DriftScore values. Because the comparison is performed in two corpora, we report Bonferroni-adjusted p-values. We also report the normalized signed-rank effect index exported by the pipeline.

Temporal transfer is evaluated by training a logistic-regression classifier on one year and testing it on another. Rows of the transfer matrix correspond to training years and columns to test years. Diagonal values are in-sample within-year references and are used only to visualize within-year separability; they should not be interpreted as out-of-sample generalization estimates.

4. RESULTS

4.1 Year-to-Year Drift in arXiv

Figure 1 compares yearly DriftScore for arXiv. The main pattern is stable across preprocessing conditions: model-level explanation drift is consistently higher than corpus-level TF–IDF and document-frequency drift. Under lemmatized preprocessing, model-level explanations have mean and median DriftScore of 0.50 across the ten year-to-year transitions, with values ranging from 0.30 to 0.60. In contrast, lemmatized TF–IDF explanations have mean and median DriftScore of 0.15, ranging from 0.00 to 0.30. Thus, model-level explanation drift is, on average, 3.33 times larger than corpus-level TF–IDF drift.

The paired Wilcoxon test confirms the separation ($p = 0.0020$; Bonferroni-adjusted $p_{adj} = 0.0040$), with a large signed-rank effect index ($r = 0.818$). This supports the central empirical finding that model-level explanation artifacts exhibit substantially greater temporal instability than corpus-level lexical summaries.

The null-distribution analysis shows that arXiv model-level drift is high but still below the conservative escalation boundary. Model-level DriftScore is above the null mean in 9 of the 10 year-to-year transitions, but it never strictly exceeds the 95th percentile of the permuted-year null distribution.

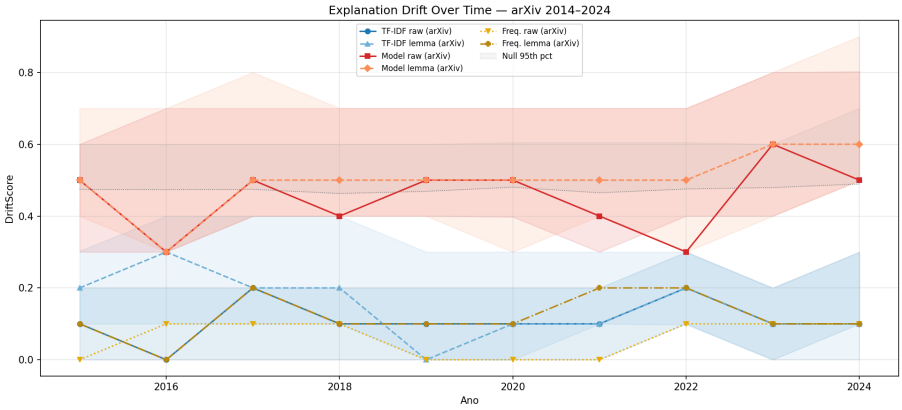

Fig. 1. Year-to-year DriftScore for arXiv. Markers and line styles distinguish TF–IDF, model-level explanations, and document-frequency baselines under raw and lemmatized preprocessing. Shaded regions show bootstrap confidence intervals and the null reference band.

In 2023, the observed value equals the null 95th percentile (0.60), but the audit rule requires a strict exceedance. In 2024, the observed value remains 0.60, while the null 95th percentile rises to 0.70. PubMed also produces no strict exceedance. Therefore, the absence of automatic audit triggers reflects the conservative calibration of the rule, not the absence of explanation turnover.

4.2 Cross-Domain Robustness Check with PubMed

Figure 2 compares model-level lemmatized drift across arXiv and PubMed. The arXiv series is consistently higher, with mean model-level DriftScore of 0.50. PubMed exhibits lower drift, with mean model-level DriftScore of 0.20 and mean TF–IDF DriftScore of 0.13.

PubMed shows the same direction with smaller magnitude. Lemmatized model-level explanations have mean and median DriftScore of 0.20, while lemmatized TF–IDF explanations have mean DriftScore of 0.13 and median DriftScore of 0.10. The model-level signal is therefore 1.54 times larger on average. However, unlike arXiv, the PubMed comparison is weaker after correction ($p = 0.0469$; $p_{adj} = 0.0938$; $r = 0.355$). This supports PubMed as a cross-domain directional robustness check rather than as a second statistically decisive benchmark.

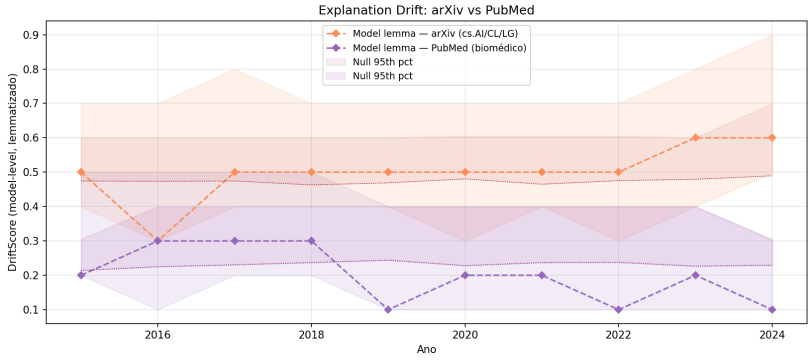

Fig. 2. Model-level explanation drift in arXiv and PubMed under lemmatized preprocessing. PubMed shows lower drift than arXiv, suggesting domain-dependent rates of explanation instability.

4.3 Pairwise Similarity and Robustness Checks

Table II illustrates the ranked terms behind the DriftScore results. TF–IDF preserves broad corpus-level terms such as *model*, *method*, *datum*, and *learning*, whereas model-level explanations replace discriminative terms more aggressively across time.

Table II. Illustrative top explanation terms in arXiv under lemmatized preprocessing.

Year	TF–IDF terms	Model-level terms
2014	algorithm, model, problem, method, language	language, word, text, datum, algorithm
2019	model, network, method, learning, datum	language, word, agent, text, image
2024	model, language, method, task, datum	language, word, model, dialogue, text

For example, the model-level top-10 sets for 2019 and 2024 share only four terms, while the corresponding TF–IDF sets share seven. Figure 3 further shows that arXiv model-level explanations are most similar near the diagonal, indicating local temporal stability and lower overlap between distant years.

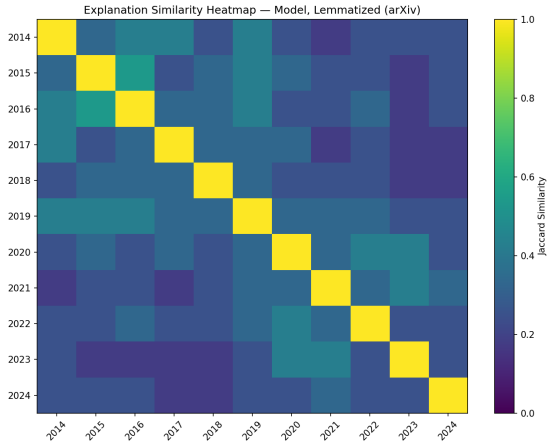

Fig. 3. Pairwise Jaccard similarity between arXiv model-level explanation artifacts under lemmatized preprocessing.

A sensitivity check with $K \in \{5, 10, 20, 30\}$ did not reverse the main conclusion: TF–IDF drift remained below model-level drift.

4.4 Temporal Transfer

The temporal-transfer matrices provide an auxiliary indication that the same period contains predictive shift. Within-year diagonal values are high, with mean accuracy of 0.943 and mean macro-F1 of 0.943, but these values are not interpreted as out-of-sample generalization estimates. Off-diagonal transfer is lower, with mean accuracy of 0.658 and mean macro-F1 of 0.657. Forward transfer also decreases with temporal distance: mean accuracy is 0.678 for one-year gaps and falls to 0.559 for the decade-shift 2014→2024 setting. This supports joint monitoring of predictive performance and explanation stability, without implying that predictive degradation directly causes explanation drift.

5. DISCUSSION

The results challenge the assumption that an interpretable mechanism necessarily provides stable explanations over time. Corpus-level drift captures changes in term salience, whereas model-level drift captures changes in how terms support classification. On arXiv, model-level drift is more than three times the mean TF–IDF drift; PubMed preserves the same direction with smaller magnitude, indicating that explanation instability is domain-dependent.

DriftScore should be interpreted descriptively: a high value indicates explanation turnover, not necessarily a need for intervention. The empirical null provides the operational distinction between *descriptive* and *actionable* drift. In our experiments, model-level drift approaches the null threshold but never produces the required strict consecutive exceedance. Therefore, the absence of alerts reflects the conservative calibration of the monitoring policy rather than the absence of explanation change.

Actionability remains application-dependent. High-risk systems may use a lower null percentile, require fewer consecutive exceedances, or combine explanation drift with predictive degradation, whereas stricter thresholds can reduce unnecessary review in lower-risk settings. The practical implication is that explanation artifacts should be versioned and monitored over time: DriftScore does not diagnose model failure, but signals when longitudinal explanation continuity has weakened.

6. LIMITATIONS AND ETHICAL CONSIDERATIONS

The study is limited to English-language scientific abstracts, which represent relatively structured textual streams. Explanation drift may behave differently in less structured and more rapidly evolving sources, such as news or social media, and in languages with different morphological, lexical, and tokenization characteristics. The arXiv corpus uses stable category labels and is therefore the primary supervised setting. PubMed relies on query-defined weak labels, so claims about biomedical classification performance must remain cautious. Its role is to test whether the monitoring behavior is directionally preserved in a different domain, not to provide a perfectly matched benchmark.

The pipeline relies on deliberately simple and interpretable mechanisms: TF–IDF, fixed vocabularies, and linear classifiers. This strengthens internal validity and makes the monitoring procedure auditable, but model-level DriftScore must also be interpreted in light of coefficient instability. NLP feature spaces contain substantial lexical redundancy and term co-occurrence, and correlated predictors may redistribute coefficient magnitude even when the underlying predictive information changes only modestly. Consequently, turnover in coefficient-based top-K explanations may reflect a combination of temporal changes in feature–label relationships and mathematical instability associated with correlated features. DriftScore should therefore be interpreted as a measure of *explanation-artifact instability*, rather than as a direct estimator of concept drift or of its underlying cause.

The monitoring architecture is not conceptually restricted to linear models. In Algorithm 1, the explanation operator E can be replaced by a post-hoc explainer applied to a fixed or explicitly versioned neural model. For example, a Transformer-based classifier could be evaluated periodically using LIME [2], SHAP [14], or Integrated Gradients [15]. Attribution magnitudes could then be aggregated across documents within each monitoring window to construct ranked explanation artifacts, after which DriftScore and empirical-null calibration could be applied without changing the core monitoring logic. Such extensions require additional methodological choices regarding local-to-global attribution aggregation, subword representation, explainer stability, computational cost, and model versioning, and are therefore left for future empirical evaluation.

Temporal-transfer results are reported only for selected off-diagonal pairs, and within-year generalization should be evaluated with proper held-out splits in future work. The current analysis also treats explanation artifacts as system outputs and does not include a user study. Future work should therefore evaluate whether DriftScore-based alerts correspond to changes that human analysts perceive as meaningful and whether alternative explanation mechanisms produce similar longitudinal behavior.

Ethically, the data are public and non-sensitive, but the deployment issue is broader. If explanations drift silently in high-stakes systems, users may infer continuity where the explanatory basis has changed. Monitoring explanation drift is therefore a governance concern, not only a technical refinement. The proposed audit rule should be understood as a configurable policy layer: stricter settings may reduce false alarms, while more sensitive settings may be necessary in applications where explanation instability has higher operational cost.

Freezing the vocabulary strengthens the deployed-monitoring interpretation because all years are evaluated through the same explanation interface. However, it may underestimate drift caused by genuinely new terminology that appears after the 2014–2016 reference window and is therefore absent from the fixed vocabulary.

7. CONCLUSION

This paper studied explanation drift in longitudinal NLP pipelines under temporal distribution shift. Using arXiv and PubMed abstracts from 2014 to 2024, we showed that model-level explanation artifacts drift more than corpus-level TF–IDF summaries, especially in the arXiv corpus. The arXiv result is statistically robust, while the PubMed result is directionally consistent and serves as a cross-domain robustness check under query-defined weak labels.

The findings support a monitoring view of explainability. Explanation reliability is not only a property of a method at one point in time; it is also a property of how its artifacts behave across time. Ranked-set overlap and DriftScore provide simple diagnostics for quantifying when explanation continuity weakens, while empirical-null calibration provides an operational mechanism for distinguishing ordinary turnover from drift that warrants review under a selected monitoring policy. The proposed audit rule is intentionally conservative: in the current data it produces no automatic triggers, but it establishes a clear boundary between descriptive explanation instability and actionable drift. In long-running NLP systems, such diagnostics should accompany predictive evaluation so that transparency remains meaningful under evolving data distributions.

REFERENCES

Doshi-Velez, F., Kim, B.: Towards a Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 (2017)

Ribeiro, M. T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: KDD (2016)

Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence **267** (2019)

Alvarez-Melis, D., Jaakkola, T. S.: On the Robustness of Interpretability Methods. ICML Workshop on Human Interpretability in Machine Learning (2018)

Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In: AIES, pp. 180–186 (2020)

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., Dennison, D.: Hidden Technical Debt in Machine Learning Systems. In: NeurIPS (2015)

Rabanser, S., Günnemann, S., Lipton, Z.: Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In: NeurIPS (2019)

Hamilton, W., Leskovec, J., Jurafsky, D.: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In: ACL (2016)

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A Survey on Concept Drift Adaptation. ACM Computing Surveys **46**(4) (2014)

Hinder, F., Vaquet, V., Hammer, B.: One or Two Things We Know About Concept Drift. Part B: Locating and Explaining Concept Drift. Frontiers in Artificial Intelligence **7** (2024)

Fumagalli, F., Muschalik, M., Hüllermeier, E., Hammer, B.: Incremental Permutation Feature Importance. Machine Learning **112**(12), 4863–4903 (2023)

Hinder, F., Vaquet, V., Brinkrolf, J., Hammer, B.: Model-based Explanations of Concept Drift. Neurocomputing **555** (2023)

Hinder, F., Vaquet, V., Hammer, B.: Feature-based Analyses of Concept Drift. Neurocomputing **600** (2024)

Lundberg, S. M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774 (2017)

Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: International Conference on Machine Learning (ICML), pp. 3319–3328 (2017)