

Article

Hypothesis-Informed Feature Stability Scoring for High-Dimensional ETL Pipelines

Konstantin Piryanov ^{1,*}, Iveta Grigorova ¹, Aleksandar Karamfilov ² and Aleksandar Efremov ¹

¹ Faculty of Automatics, Industrial Automation, Technical University of Sofia, 1000 Sofia, Bulgaria; iveta.grigorova@tu-sofia.bg (I.G.); alefremov@tu-sofia.bg (A.E.)

² Faculty of German Education in Engineering and Business Administration, Technical University of Sofia, 1000 Sofia, Bulgaria; akaramfilov@tu-sofia.bg

* Correspondence: konstantin.piryanov@tu-sofia.bg

Abstract

High-dimensional financial Extract–Transform–Load (ETL) pipelines often contain heterogeneous variables whose statistical properties may change between recurring data deliveries, affecting feature reliability before downstream machine learning models are trained. This study extends a previously proposed Canberra-based data drift monitoring framework by introducing a hypothesis-informed feature stability component for automated feature assessment and prioritization. Unlike the prior descriptive framework, which relied on univariate and bivariate exploratory metrics, the proposed extension adds an inferential layer and evaluates how this layer changes feature ranking relative to the original score and alternative marginal drift measures. The method combines univariate deviations in summary statistics, bivariate deviations in dependency-related metrics, and hypothesis-based evidence from Anderson–Darling, Mann–Whitney U, and Levene tests. The resulting *p*-values are aggregated using a Landau-calibrated harmonic mean *p*-value formulation and transformed into a bounded hypothesis score, which is integrated into a composite variable-level stability ranking. The framework operates on precomputed exploratory data analysis (EDA) outputs, enabling scalable comparison between a validated reference dataset and a current ETL delivery. The proposed extension provides an interpretable and computationally efficient mechanism for identifying unstable features and supporting feature review, exclusion, or prioritization in automated machine learning pipelines.

Keywords: feature stability; feature selection; data drift; hypothesis testing; harmonic mean *p*-value; ETL monitoring; Canberra distance; high-dimensional data

Academic Editor: Rui Araújo

Received: 20 May 2026

Revised: 21 June 2026

Accepted: 26 June 2026

Published: 28 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The rapid growth of data storage, computational capacity, and automated analytics has enabled organizations to process increasingly large and heterogeneous datasets. In many applied machine learning environments, especially in financial analytics, this growth has produced high-dimensional data pipelines containing hundreds or thousands of candidate variables. While such feature-rich datasets can improve predictive modeling, they also introduce practical challenges related to computational complexity, duplicated information, unstable variables, and the curse of dimensionality. Feature extraction and feature selection methods have therefore become essential tools for reducing model

complexity, improving interpretability, and supporting efficient model development [1–4]. Recent survey work also confirms that feature selection remains an active topic in machine learning, with continuing emphasis on reducing dimensionality, addressing redundancy, and selecting methods appropriate to dataset characteristics and modeling objectives [5].

Classical feature selection methods are commonly grouped into filter, wrapper, embedded, and hybrid approaches. These methods typically evaluate features according to predictive relevance, redundancy, statistical association with a target variable, or contribution to model performance [1–4]. Such approaches are valuable when the training dataset is assumed to be representative of future data. However, in recurring production environments, this assumption may not hold. Features that appear predictive during one modeling cycle may become unreliable in later cycles because of upstream changes in data definitions, source systems, transformations, missingness patterns, or population composition. As a result, feature selection should not only consider predictive utility, but also the temporal and statistical stability of candidate variables. This view is consistent with research on feature-selection stability, where instability in selected feature subsets is treated as an important concern for reproducible and reliable modeling [6].

This issue is particularly important in recurring Extract–Transform–Load (ETL) workflows used for automated financial modeling. In such systems, new data deliveries are often received periodically from multiple institutions or operational sources. Even when schema validation checks pass, the statistical behavior of individual variables may change substantially between deliveries. A variable may retain the same name and data type while its distribution, cardinality, missingness structure, or relationship with other variables changes. These changes can silently affect downstream machine learning models, leading to unstable feature importance, degraded model reliability, or inconsistent decision-making. Existing data validation systems are effective for detecting missing columns, invalid types, or rule violations, but they are less suitable for identifying subtle distributional and relational drift in high-dimensional heterogeneous data [7–9]. Recent work on data quality requirements for machine learning development pipelines further shows that data-quality concerns affect multiple lifecycle stages, including data collection, preprocessing, training, validation, deployment, and monitoring [10]. Related activity-monitoring research has also emphasized the importance of detecting meaningful changes in evolving populations before they become visible through downstream outcomes [11].

A related body of research addresses concept drift, dataset shift, and production machine learning monitoring. Concept drift studies generally focus on changes in the relationship between input variables and target outcomes over time [12–14]. Dataset shift methods focus on changes in the input or joint data distribution [15–17]. Recent benchmark and survey work on fully unsupervised concept drift detection also highlights the practical importance of detecting distributional changes when labels are delayed, unavailable, or expensive to obtain [18]. Distributional drift detection methods often rely on statistical distances or divergence measures such as Kullback–Leibler divergence and Jensen–Shannon divergence [19,20], Wasserstein distance [21,22], Kolmogorov–Smirnov statistics [23], or Population Stability Index. These approaches are useful but can be difficult to apply uniformly in operational ETL settings involving mixed numerical, categorical, and ordinal variables. Some methods require full empirical distributions, density estimation, binning decisions, or raw-data access, which may increase computational cost and reduce interpretability in large recurring pipelines.

In previous work, a Canberra-based ETL drift monitoring framework was proposed to address this gap by comparing a validated reference dataset with a current ETL delivery using precomputed exploratory data analysis (EDA) metrics [24]. The framework aggregated univariate and bivariate summary-statistic deviations into bounded variable-

level stability scores, allowing variables to be ranked by drift severity without requiring trained model artifacts or full probability density reconstruction. The original study demonstrated that the score responded monotonically to controlled noise and could identify schema evolution, shifts in missingness, and disruptions in dependency structures. The method used a hierarchical aggregation strategy in which univariate and bivariate sub-scores contributed equally to the final variable-level score, preserving interpretability and reducing the risk that one statistical dimension would dominate the ranking.

Despite these advantages, the original framework was primarily descriptive. It quantified the magnitude of statistical and relational deviations but did not incorporate formal inferential evidence into the final scoring logic. Therefore, the previous score could rank variables according to observed EDA-based deviations, but it could not indicate whether these deviations were supported by hypothesis-test evidence. The present study addresses this limitation by adding a hypothesis-informed component, evaluating alternative p -value aggregation methods, and examining how the resulting inferential signal changes feature prioritization relative to the original Canberra-based score. Hypothesis-based validation, including distributional, location, and dispersion tests, was identified as a future extension that would help distinguish random deviations from systematic drift [24]. Similarly, comparison with alternative drift measures such as Jensen–Shannon divergence and Wasserstein distance was left for future work as a way to provide a more complete characterization of distributional change [24]. These open directions motivate the present study.

This perspective is also consistent with recent data-centric artificial intelligence research, which emphasizes that reliable machine learning systems require systematic attention to data development, inference data monitoring, and data maintenance, rather than model optimization alone [25].

The present work extends the Canberra-based framework by introducing a hypothesis-informed feature stability component. For each aligned numerical variable, the proposed extension computes complementary two-sample statistical tests: the Anderson–Darling test to assess distributional shape differences, the Mann–Whitney U test to detect rank-based location shifts, and Levene’s test to evaluate changes in dispersion. The resulting p -values are combined using alternative p -value aggregation strategies, with particular attention to the harmonic mean p -value (HMP), which has been proposed as a method for combining dependent statistical evidence while maintaining useful power properties in large-scale testing settings [26]. The combined p -value is then transformed into a bounded hypothesis evidence score and integrated with the existing univariate and bivariate Canberra-based scores.

The aim of this study is therefore to develop and evaluate a hypothesis-informed feature stability ranking method for high-dimensional ETL-based machine learning pipelines. Specifically, the study pursues four objectives: first, to extend the existing Canberra-based stability score with inferential evidence from statistical hypothesis tests; second, to compare alternative p -value combination methods and justify the use of the harmonic mean p -value in this context; third, to compare the resulting composite score with established distributional drift measures, including Jensen–Shannon and Wasserstein-based alternatives; and fourth, to interpret feature stability ranking as a pre-modeling feature assessment mechanism that can support feature review, exclusion, or prioritization.

The novelty of the present work lies in transforming the original descriptive Canberra-based drift score into a hypothesis-informed stability score by integrating test-based inferential evidence, p -value aggregation, and comparative rank evaluation against both the original baseline and established marginal drift measures. The main contribution of this work is the treatment of feature stability as a complementary dimension of feature selection in automated machine learning pipelines. Unlike conventional feature selection,

which primarily evaluates predictive relevance or redundancy, the proposed approach evaluates whether candidate variables remain statistically reliable across recurring data deliveries. By combining descriptive deviation, dependency change, and hypothesis-based evidence, the framework provides an interpretable and computationally efficient mechanism for identifying unstable features before model training. In this way, the proposed method contributes to machine learning-based feature assessment by linking data drift monitoring with feature selection support in high-dimensional, production-oriented environments.

2. Materials and Methods

This study extends the previously proposed Canberra-based ETL drift monitoring framework by introducing an inferential feature stability component and evaluating its behavior against alternative p -value aggregation and distributional drift measures. The original framework compared a validated reference dataset with a current ETL delivery using precomputed exploratory data analysis (EDA) metrics and produced variable-level stability scores based on univariate and bivariate deviations. In the present work, this baseline formulation is retained and expanded in three directions: first, by incorporating hypothesis-based evidence derived from statistical tests; second, by comparing alternative methods for combining p -values into a single variable-level inferential signal; and third, by evaluating the resulting composite score against established distributional drift measures, including Jensen–Shannon divergence and Wasserstein distance.

Let D_{ref} denote the validated reference dataset and D_t denote the current ETL delivery at time t . For each aligned feature v , the framework computes three complementary stability components that are combined into a final feature stability ranking. Higher scores indicate stronger evidence of feature instability and therefore higher priority for review before downstream machine learning model training.

The proposed methodology is intended as a pre-modeling feature assessment layer. It does not replace conventional feature selection methods based on predictive relevance, redundancy, or model performance. Instead, it provides a complementary stability-oriented ranking for identifying variables that may be unreliable across recurring data deliveries. Such variables may require exclusion, transformation, additional validation, or reduced modeling priority.

Due to confidentiality restrictions, raw observations, institution-specific information, and original feature names are not disclosed. All variables are anonymized using generic identifiers before reporting. The anonymization preserves feature type, score values, ranks, and comparative evaluation outputs, but removes semantic labels and business-specific information. Section 3 therefore reports score distributions, anonymized feature rankings, and rank-agreement statistics rather than raw feature distributions or identifiable variable names.

2.1. Study Design and Framework Overview

The study is designed as a methodological extension and comparative evaluation of a previously developed feature drift monitoring framework. The workflow consists of five stages: reference-current data alignment, baseline Canberra-based scoring, hypothesis-based testing, p -value aggregation, and comparison with alternative drift measures.

Figure 1 summarizes the proposed methodology. The workflow begins with alignment between the validated reference dataset and the current ETL delivery. Aligned features are then evaluated using two complementary branches: the baseline Canberra-based branch, which computes univariate and bivariate stability components from EDA summary metrics, and the hypothesis-based branch, which applies two-sample tests to stored feature samples and aggregates the resulting p -values using the HMP procedure. The

resulting components are combined into the final composite feature stability score used for ranking and pre-modeling review.

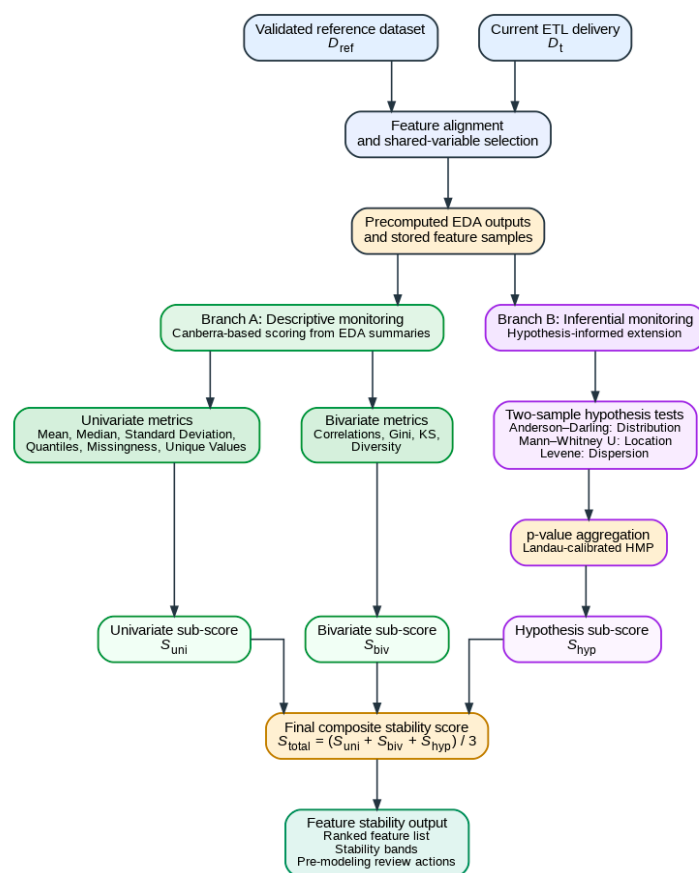


Figure 1. Workflow of the proposed hypothesis-informed feature stability scoring framework. The retained Canberra-based branch computes S_{uni} and S_{biv} from EDA summary metrics, while the proposed hypothesis-informed branch computes S_{hyp} from Anderson–Darling, Mann–Whitney U, and Levene tests followed by harmonic mean p -value aggregation. The three components are combined into S_{total} , which supports feature ranking, stability-band assignment, and pre-modeling review. Source: Authors’ own elaboration.

First, schema alignment is performed between D_{ref} and D_t . Variables present in both datasets are retained as the aligned feature set for feature stability scoring, while added or removed variables are reported separately as schema differences. This ensures that the composite score is computed only for comparable features, while structural changes in the dataset are still made visible in diagnostic output.

Before scoring, the reference dataset was treated as the validated baseline ETL output. Data reliability and validity were assessed through schema alignment, type consistency checks, missingness profiling, uniqueness and cardinality checks, range and distributional summaries, and dependency-related exploratory data analysis metrics. Features not present in both the reference and current datasets were recorded as schema differences and excluded from composite scoring. This procedure ensured that the stability score was computed only for comparable variables, while structural ETL changes were preserved as separate diagnostic information.

Second, the retained variables are evaluated using the baseline Canberra-based framework. This component produces two scores for each feature: a univariate score, S_{uni} , based on deviations in marginal summary statistics, and a bivariate score, S_{biv} , based on

deviations in dependency-related metrics. These scores quantify descriptive and relational feature instability.

Third, a hypothesis-based component is introduced. For each eligible numerical feature, statistical tests are applied to compare the reference and current samples. The tests are selected to capture complementary types of change: distributional shape differences, location shifts, and dispersion changes.

Fourth, the resulting p -values are aggregated into a single feature-level inferential signal. Several p -value combination methods are compared, with the harmonic mean p -value used as the primary aggregation strategy.

Finally, the proposed composite score is compared with established distributional drift measures. This comparison evaluates whether the proposed score produces stable and interpretable feature rankings while remaining compatible with high-dimensional ETL outputs.

The overall workflow is summarized as follows:

$$D_{ref}, D_t \rightarrow \{S_{uni}, S_{biv} \mid p_{AD}, p_{MW}, p_{LEV} \rightarrow S_{hyp} \rightarrow S_{total}, \quad (1)$$

where D_{ref} denotes the validated reference dataset; D_t denotes the current ETL delivery at time t ; S_{uni} and S_{biv} are the univariate and bivariate Canberra-based stability components inherited from the baseline framework; p_{AD} , p_{MW} and p_{LEV} are the p -values produced by the Anderson–Darling, Mann–Whitney U, and Levene tests, respectively; S_{hyp} is the newly introduced hypothesis-informed score obtained from p -value aggregation; and S_{total} is the final composite feature stability score.

2.2. Baseline Canberra-Based Feature Stability Score

The baseline component of the proposed method follows the Canberra-based scoring framework introduced in [24]. The original framework was designed to compare recurring ETL outputs using summary statistics rather than full raw-data distributions. This design allows the monitoring layer to operate on precomputed exploratory data analysis outputs, reducing computational overhead while preserving interpretability.

For each monitored feature v , the associated metrics are divided into two groups: univariate metrics M_{uni} and bivariate metrics M_{biv} . Univariate metrics include the number of unique values, missing-value proportion, mean, median, standard deviation, and selected quantiles. Bivariate metrics include correlation and dependency- or distribution-related indicators.

For a given metric m , let x^r and x^c denote the metric values computed from the reference dataset D_{ref} and the current dataset D_t , respectively. Selected metrics are normalized before aggregation to ensure comparability across heterogeneous scales. In particular, the number of unique values is normalized using the joint reference-current mass, and percentage-based bivariate metrics such as Gini, Diversity, and Kolmogorov–Smirnov statistics are divided by 100 before aggregation.

The component-wise Canberra distance is defined as:

$$d_m(\hat{x}^r, \hat{x}^c) = \frac{|\hat{x}^r - \hat{x}^c|}{|\hat{x}^r| + |\hat{x}^c| + \epsilon}, \quad (2)$$

where $d_m(\hat{x}^r, \hat{x}^c)$ denotes the Canberra deviations for metric m ; $\hat{x}^r$ and $\hat{x}^c$ are the normalized reference and current metric values, respectively; and $\epsilon = 10^{-12}$ is a numerical stabilization constant used to avoid division by zero.

The univariate and bivariate scores are computed as the average Canberra deviations over their respective metric groups:

$$S_{uni} = \frac{1}{|M_{uni}|} \sum_{m \in M_{uni}} d_m, \quad S_{biv} = \frac{1}{|M_{biv}|} \sum_{m \in M_{biv}} d_m. \quad (3)$$

In the original framework, the baseline variable-level stability score was computed as:

$$S_{base} = \frac{S_{uni} + S_{biv}}{2}, \quad (4)$$

where S_{base} denotes the original two-component Canberra-based feature stability score, computed as the average of the univariate S_{uni} and the bivariate component S_{biv} . This score represents the baseline formulation from the previous framework before adding the hypothesis-informed component.

This hierarchical aggregation ensures that marginal statistical properties and dependency-related structure contribute equally to the baseline feature stability score. The present study retains S_{uni} and S_{biv} as the descriptive foundation of the framework and extends the scoring logic by introducing the hypothesis-based component S_{hyp} .

2.3. Hypothesis-Based Feature Stability Component

The baseline Canberra-based score quantifies descriptive deviations between reference and current exploratory data analysis metrics. However, descriptive differences alone do not indicate whether the observed changes are statistically supported or may plausibly arise from random sampling variation. To address this limitation, the present study introduces a hypothesis-based feature stability component, denoted as S_{hyp} .

For each aligned numerical feature v , let X_v^r denote the sample of observed values in the reference dataset D_{ref} , and let X_v^c denote the corresponding sample in the current dataset D_t . The tests are interpreted as population-level comparisons between ETL deliveries rather than as paired record-level tests, because individual entities may partially overlap between the reference and current datasets. Before testing, missing values, positive infinity, and negative infinity are removed. Features with no valid observations after filtering are excluded from hypothesis testing. When a test cannot be computed because of constant values or insufficient variability, its p -value is recorded as missing and excluded from p -value aggregation.

The hypothesis component uses three complementary two-sample tests. The Anderson–Darling k -sample test is used to detect differences in distributional shape and tail behavior [27]. The Mann–Whitney U test is used to detect rank-based location shifts without assuming normality [28]. Levene’s test with median centering is used to detect changes in dispersion and is more robust to non-normality than classical variance tests based on deviations from the mean [29,30].

These three tests were selected because they provide complementary and interpretable evidence of feature-level instability: Anderson–Darling captures distributional shape and tail differences, Mann–Whitney U captures rank-based location shifts, and Levene’s test captures dispersion changes. The selected tests therefore cover three common forms of numerical feature instability while keeping the hypothesis component compact and interpretable.

For diagnostic interpretation, the main hypothesis signal was assigned according to the valid test with the smallest p -value for each feature. Anderson–Darling was labeled as “Distribution”, Mann–Whitney U as “Location”, and Levene’s test as “Dispersion”. This label was used only to support the interpretation of the dominant hypothesis-test evidence and did not affect the harmonic mean p -value aggregation, S_{hyp} or the final composite score S_{total} .

For each feature, the available p -values are denoted as:

$$P_v = \{p_{AD}, p_{MW}, p_{LEV}\}, \quad (5)$$

where missing or non-applicable p -values are excluded before aggregation. These p -values should not be treated as fully independent because all tests are applied to the same pair of samples and may respond to related forms of distributional change. Therefore, the individual p -values are combined into a single inferential signal using p -value aggregation methods.

The hypothesis score is interpreted as evidence of instability rather than as a measure of drift magnitude. This distinction is important because p -values are sensitive to sample size: small effects may become statistically significant in large datasets, while larger descriptive changes may not be significant in smaller or sparse samples. For this reason, S_{hyp} is used as a complementary component rather than as a replacement for the Canberra-based descriptive scores.

Because the hypothesis component is computed from stored feature samples generated during the exploratory data analysis stage, sampling variability may affect individual p -values and therefore the resulting S_{hyp} ranking, especially for features with similar scores. To reduce uncontrolled variability and ensure reproducibility, sampling was performed using a fixed random state. In addition, missing and non-finite values were removed before testing, and unequal sample sizes were allowed for the Anderson–Darling, Mann–Whitney U, and Levene tests. Accordingly, S_{hyp} should be interpreted as a reproducible sample-based inferential indicator rather than an exact full-population hypothesis-test result.

2.4. p -Value Combination Methods

The hypothesis-based component produces up to three p -values for each monitored numerical feature. Because each test captures a different form of instability, a single combined inferential signal is required for feature-level ranking.

$$P_v = \{p_1, p_2, \dots, p_k\}, \quad (6)$$

The set of valid p -values available for each feature is denoted as v , where $1 \leq k \leq 3$. Invalid or missing p -values are excluded, and valid p -values are constrained to the interval $(0,1]$ using a small numerical constant to avoid numerical instability. In this study, the full planned test family size is $L = 3$, corresponding to the Anderson–Darling, Mann–Whitney U, and Levene tests.

Several p -value combination methods were evaluated. Their roles in the study are summarized in Table 1.

Table 1. p -value combination methods used for comparison. Source: Authors' own elaboration.

Method	Aggregation Principle	Expected Behavior in Feature Stability Ranking
Fisher's method [31]	Combines logarithmic evidence across p -values	Sensitive to several moderately small p -values
Stouffer's method [32]	Combines normal-transformed p -values	Useful for averaged evidence and possible weighting
Tippett's method [33]	Uses the smallest p -value	Sensitive to one strong test signal
Bonferroni-type minimum- p adjustment [34]	Applies conservative correction to the smallest p -value	Reduces false positives but may reduce sensitivity
Harmonic mean p -value [26]	Combines p -values through their harmonic mean	Balances sensitivity to small p -values with information from all tests

The primary aggregation strategy used in the proposed framework is based on the harmonic mean p -value [26]. For the valid p -values P_v , the raw harmonic mean p -value statistic is computed as:

$$p_H = \frac{k}{\sum_{i=1}^k 1/p_i}. \quad (7)$$

The harmonic mean formulation is sensitive to small p -values while still incorporating information from all available tests. This is appropriate in the present setting because distributional, location, and dispersion tests provide related but non-identical evidence of feature instability.

Because the p -values are obtained from tests applied to the same pair of samples, they should not be assumed to be fully independent. Therefore, a Landau-calibrated HMP-style procedure is used as an inferential evidence score for feature ranking, not as a standalone formal multiple-testing decision rule. The Landau-calibrated HMP-style score was selected because the individual p -values are produced from tests applied to the same reference-current feature samples and may therefore represent dependent evidence. In this setting, Fisher's and Stouffer's methods are useful comparators, but their classical interpretations are more naturally associated with independence or explicit weighting assumptions. Tippett's and Bonferroni-type procedures emphasize the smallest p -value, which is useful when one test provides a strong signal but may underuse supporting evidence from the remaining tests. The HMP formulation provides a compromise: it remains sensitive to small p -values while incorporating all available valid tests. For this reason, the Landau-calibrated HMP-style score is used as the primary inferential ranking component, while Fisher's, Stouffer's, Tippett's, and Bonferroni-type aggregation are retained for comparative analysis. Let p_H denote the raw harmonic mean p -value statistic. The calibration is computed using:

$$x = \frac{1}{p_H}. \quad (8)$$

$$p_L = P(\text{Landau}(\log(L) + 0.874, \pi/2) \geq x), \quad (9)$$

where p_L is the Landau-calibrated tail probability. To avoid overstating evidence from the raw harmonic mean statistic, the combined p -value is defined conservatively as:

$$p_{combined} = \max(p_H, p_L). \quad (10)$$

The combined p -value is then transformed into a bounded evidence score:

$$S_{hyp} = 1 - p_{combined}. \quad (11)$$

Higher values of S_{hyp} therefore indicate stronger inferential evidence of feature instability. The HMP-based score is used for feature ranking and diagnostic prioritization rather than as a standalone formal multiple-testing decision rule.

The p -value combination methods are compared in Section 3 using rank-based criteria, including Spearman rank correlation [35], Kendall rank correlation [36], and top- k overlap among the highest-ranked unstable features.

2.5. Composite Feature Stability Score

After computing the baseline Canberra-based components and the hypothesis-based component, the three signals are integrated into a single composite feature stability score.

For each aligned feature v , the available component scores are:

- $S_{uni}(v)$, capturing marginal or univariate deviation;
- $S_{biv}(v)$, capturing dependency-related or bivariate deviation;

- $S_{hyp}(v)$, capturing inferential evidence of distributional, location, or dispersion change.

When all three components are available, the final composite score is defined as:

$$S_{total}(v) = \frac{S_{uni}(v) + S_{biv}(v) + S_{hyp}(v)}{3}, \quad (12)$$

where $S_{total}(v)$ denotes the final hypothesis-informed composite feature stability score for feature v , and $S_{uni}(v)$, $S_{biv}(v)$ and $S_{hyp}(v)$ denote univariate, bivariate, and hypothesis-informed score components for the same feature, respectively.

Higher values indicate stronger instability. Features with low S_{total} are interpreted as stable relative to the reference dataset, while features with high S_{total} are prioritized for review before downstream model training.

If the hypothesis component is unavailable for a feature, for example because the variable is non-numeric or because no valid hypothesis-test p -values can be computed, the composite score is calculated as the mean of the available components:

$$S_{total}(v) = \frac{\sum_{q \in Q_v} S_q(v)}{|Q_v|}, \quad (13)$$

where Q_v denotes the set of score components available for feature v . $S_q(v)$ is the value of component q for that feature, and $|Q_v|$ is the number of available components. In the complete case, $Q_v = \{S_{uni}, S_{biv}, S_{hyp}\}$. If the hypothesis component is unavailable, for example because no valid hypothesis-test p -value can be computed, the score is calculated as the average of the remaining available components. This formulation preserves comparability while retaining information about component availability in the diagnostic output.

Equal weighting is used deliberately. The goal is not to optimize feature selection for a specific predictive model, but to provide a transparent, model-agnostic stability ranking. Assigning equal weight to the univariate, bivariate, and hypothesis components ensures that marginal distributional changes, dependency-structure changes, and inferential evidence each contribute to the final ranking.

For interpretation, the dominant instability signal is assigned according to the largest available component score:

$$Signal(v) = \operatorname{argmax}_{q \in \{uni, biv, hyp\}} S_q(v), \quad (14)$$

where $Signal(v)$ denotes the dominant instability signal assigned to feature v , and argmax returns the available component q with the largest score $S_q(v)$. The resulting label is used only as a diagnostic aid and does not modify the final composite score. It indicates whether the feature was prioritized mainly because of marginal change, dependency-structure change, or hypothesis-supported evidence of instability.

The resulting S_{total} values are used to produce a ranked list of the features. This ranking supports pre-modeling feature assessment by identifying variables that may require validation, transformation, exclusion, or monitoring before being used in downstream machine learning algorithms.

2.6. Alternative Distributional Drift Measures for Comparison

To evaluate the behavior of the proposed composite feature stability score, two established distributional drift measures are used as comparative baselines: Jensen–Shannon divergence [20] and Wasserstein distance [21,22]. These measures are selected because they are widely used for comparing empirical distributions and represent different perspectives on distributional change.

The purpose of this comparison is not to replace the proposed composite score, but to assess whether the ranking produced by S_{total} is consistent with conventional

distributional drift indicators. Unlike the proposed score, which integrates univariate deviations, bivariate deviations, and hypothesis-based evidence, Jensen–Shannon divergence and Wasserstein distance primarily compare marginal feature distributions.

The Canberra-based S_{uni} and S_{biv} components are computed from full EDA summary metrics. In contrast, S_{hyp} , Jensen–Shannon divergence, and Wasserstein distance are computed from stored feature samples generated during the exploratory data analysis stage. Therefore, the alternative distributional measures are interpreted as sample-based marginal drift indicators used for rank-based comparison rather than exact full-dataset divergence estimates. Consequently, ranking differences involving S_{hyp} , Jensen–Shannon divergence, and Wasserstein distance may partly reflect sampling variability, particularly when feature scores are close.

For Jensen–Shannon divergence, the empirical reference and current distributions of feature v are denoted as P_v and Q_v . The midpoint distribution is:

$$M_v = \frac{1}{2}(P_v + Q_v). \quad (15)$$

The Jensen–Shannon divergence is computed as:

$$JSD(P_v, Q_v) = \frac{1}{2}KL(P_v \parallel M_v) + \frac{1}{2}KL(Q_v \parallel M_v), \quad (16)$$

where $JSD(P_v, Q_v)$ denotes the Jensen–Shannon divergence between the empirical reference distribution P_v and the empirical current distribution Q_v of feature v , M_v is the midpoint distribution defined in Equation (15), and $KL(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence. For numerical variables, empirical distributions are obtained using a common binning scheme applied to both reference and current samples. For categorical and ordinal variables, normalized frequency tables are used. A small smoothing constant is applied to avoid undefined logarithmic terms caused by zero probabilities.

Wasserstein distance is used as a geometry-based measure of distributional displacement. For numerical variables, the first Wasserstein distance compares the empirical reference sample X_v^r and the current sample X_v^c . Because Wasserstein distance is expressed in the units of the feature, it is normalized to support comparison across heterogeneous variables:

$$W_{norm}(v) = \frac{W_1(X_v^r, X_v^c)}{IQR(X_v^r \cup X_v^c) + \epsilon}, \quad (17)$$

where $W_{norm}(v)$ denotes the normalized Wasserstein distance for feature v , $W_1(X_v^r, X_v^c)$ is the first Wasserstein distance between the reference sample X_v^r and current sample X_v^c , $IQR(X_v^r \cup X_v^c)$ is the interquartile range of the combined reference-current sample, and $\epsilon = 10^{-12}$ is a numerical stabilization constant used to avoid division by zero. Wasserstein distance is applied only to numerical and meaningfully ordered ordinal variables. It is not used as a primary comparator for purely nominal categorical variables because category codes do not necessarily represent meaningful geometric distance.

The alternative drift measures are compared with S_{total} using rank-based criteria rather than direct score equality, because the methods operate on different mathematical scales and capture different aspects of feature instability. The comparison uses Spearman rank correlation, Kendall rank correlation, and top-k overlap:

$$Overlap_k(A, B) = \frac{|Top_k(A) \cap Top_k(B)|}{k}, \quad (18)$$

where $Top_k(A)$ and $Top_k(B)$ are the sets of the k highest-ranked features according to two different scoring methods.

Disagreement between S_{total} and alternative measures is not necessarily considered an error. Jensen–Shannon divergence and Wasserstein distance focus mainly on marginal

distributional change, whereas the proposed score also incorporates dependency-structure deviations and hypothesis-based evidence. Therefore, component-level diagnostics are used to interpret cases where the composite ranking differs from the marginal distributional measures.

All computations were implemented in Python 3.11.15 (Python Software Foundation, Wilmington, DE, USA) using NumPy 2.0.2, pandas 2.3.1, and SciPy 1.15.0. The same aligned feature set was used across the Canberra-based score, hypothesis component, and comparative distributional measures where methodologically applicable. Numerical stabilization constants were applied consistently to avoid division-by-zero, logarithmic, and p -value boundary artifacts. When Landau-calibrated HMP-style scoring was used, the SciPy environment was required to support the corresponding Landau distribution implementation.

3. Results

The proposed scoring framework was applied to the aligned feature set obtained after matching the validated reference dataset D_{ref} and the current ETL delivery D_t . After schema alignment, 326 variables were retained for composite feature stability scoring. Due to confidentiality restrictions, all feature names are anonymized and reported using generic identifiers. The results therefore focus on score distributions, component-level contributions, feature ranking, and rank-agreement statistics rather than raw observations or institution-specific feature semantics. This section first summarizes the distribution of feature stability scores across all aligned variables, then examines the highest- and lowest-ranked features, evaluates the hypothesis-based component and p -value aggregation methods, compares the composite score with baseline and alternative drift measures, and finally discusses implications for pre-modeling feature selection.

3.1. Distribution of Feature Stability Scores

Table 2 summarizes the distribution of the three component scores and the final composite score across the 326 aligned variables. The univariate component S_{uni} has a low mean value of 0.0391 and a median of 0.0161, indicating that most aligned variables show limited marginal deviation in summary statistics. However, the maximum value of 0.8551 shows that a small subset of variables exhibits substantial univariate instability.

The bivariate component S_{biv} is more dispersed, with a mean of 0.2203 and a median of 0.1559. This indicates that dependency-related changes are more pronounced than marginal summary-statistic changes for a larger proportion of variables. The maximum value of 1.0000 shows that at least one feature exhibited the strongest possible bivariate instability under the scoring formulation.

The hypothesis component S_{hyp} has the highest overall values, with a mean of 0.8262 and a median of 0.9972. This suggests that many variables show strong inferential evidence of change between the reference and current samples. This result should be interpreted together with the descriptive components because S_{hyp} reflects statistical evidence rather than drift magnitude. In large datasets, even relatively small distributional, location, or dispersion changes can produce strong hypothesis-test evidence.

The final composite score S_{total} has a mean of 0.3613 and a median of 0.3670, indicating moderate overall feature instability across the aligned feature set. The interquartile range, from 0.3172 to 0.4201, shows that most variables are concentrated within a relatively narrow middle band, while the maximum value of 0.9514 indicates that a limited number of variables are strongly prioritized for review.

Table 2. Summary statistics of feature stability scores across aligned anonymized variables. Source: Authors' own elaboration.

Score	Mean	Median	Std. Dev.	Min	Q1	Q3	Max
S_{uni}	0.0391	0.0161	0.0958	0.0000	0.0049	0.0266	0.8551
S_{biv}	0.2203	0.1559	0.2189	0.0000	0.0467	0.3202	1.0000
S_{hyp}	0.8262	0.9972	0.2691	0.0000	0.6737	0.9999	1.0000
S_{total}	0.3613	0.3670	0.1313	0.0000	0.3172	0.4201	0.9514

Figure 2 presents the distribution of the final composite feature stability score S_{total} . The score distribution provides an overview of instability across the aligned anonymized variables and identifies whether instability is concentrated in a small subset of features or distributed more broadly across the feature space.

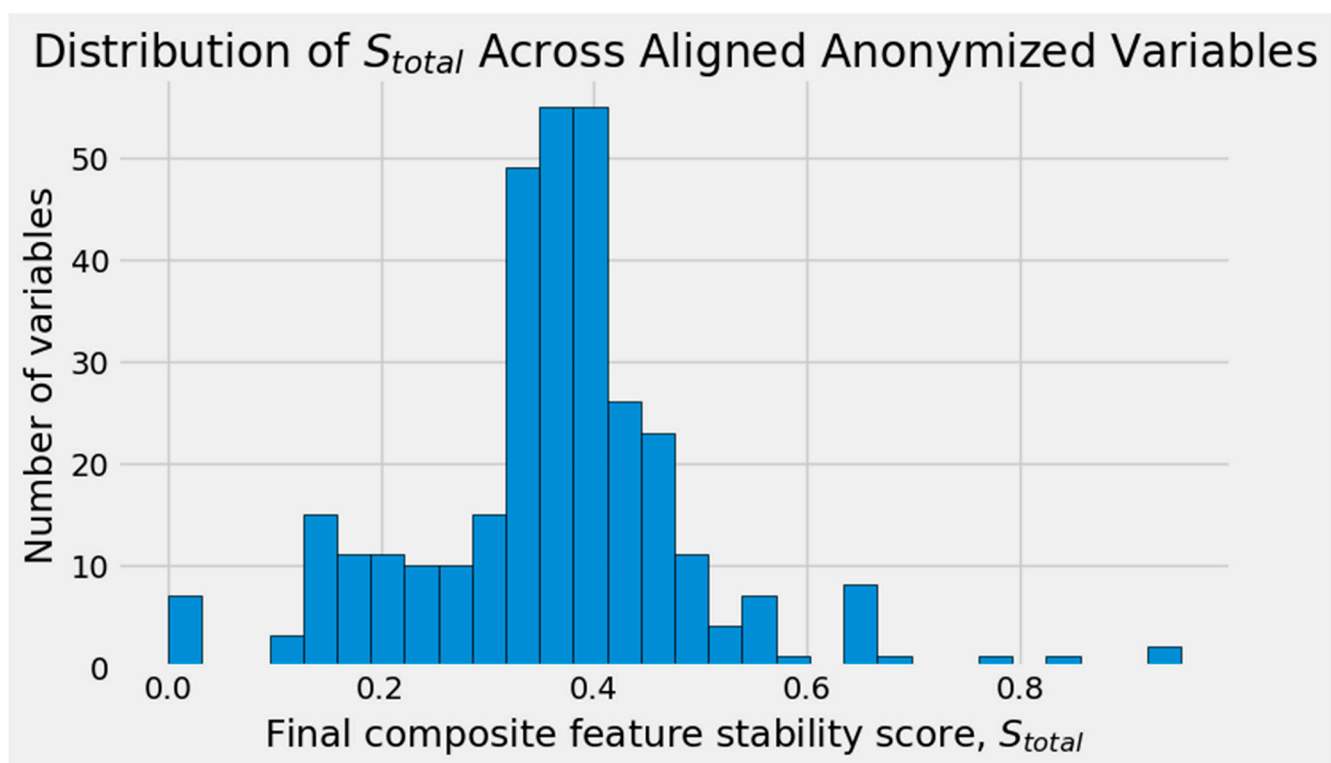


Figure 2. Distribution of the final composite feature stability score S_{total} across aligned anonymized variables. Source: Authors' own elaboration.

Figure 3 compares the distributions of S_{uni} , S_{biv} , S_{hyp} and S_{total} . The component-wise comparison shows that the hypothesis component contributes the strongest evidence across the feature set, while the univariate component remains low for most variables. This indicates that many detected changes are statistically supported but not necessarily large in marginal magnitude. The bivariate component provides an intermediate signal, capturing relational changes that are not fully represented by univariate summary statistics.

Overall, the score distributions indicate that the proposed framework separates three complementary dimensions of feature instability: limited marginal deviation for most variables, more visible dependency-related variation, and strong hypothesis-supported evidence of change. The composite score moderates these signals into a single ranking suitable for prioritizing variables for further review before downstream model training.

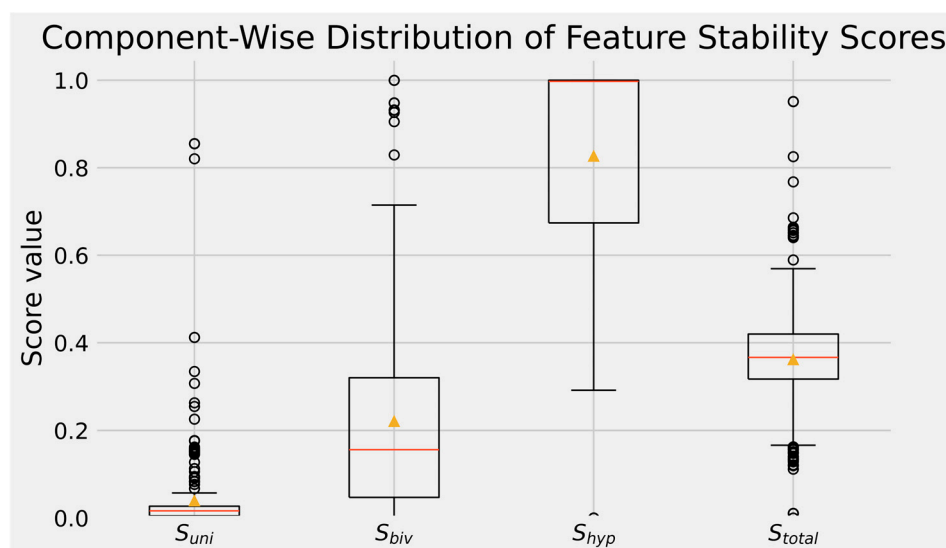


Figure 3. Component-wise distribution of S_{uni} , S_{biv} , S_{hyp} , and S_{total} across aligned anonymized variables. Boxplot elements follow standard conventions: boxes indicate interquartile range, whiskers indicate the non-outlier range, red horizontal lines indicate medians, yellow triangles indicate means, and black circles indicate outliers. Source: Authors' own elaboration.

3.2. Highest- and Lowest-Ranked Anonymized Features

The final composite score S_{total} was used to rank the aligned variables according to overall feature stability. Tables 3 and 4 report the ten highest- and lowest-ranked variables, respectively. To preserve confidentiality, variables are reported using anonymized identifiers. The component scores are shown to indicate whether each feature was prioritized mainly because of marginal statistical deviation, dependency-structure deviation, or hypothesis-supported evidence of change.

Table 3 presents the ten highest-ranked variables. The two most unstable features achieved final scores above 0.95, with both showing very high univariate, bivariate, and hypothesis-based components. Their dominant signal was bivariate, indicating that changes in dependency-related behavior were the largest contributors to their final ranking. The third-ranked feature also showed high overall instability, but its dominant signal was hypothesis-based, with $S_{hyp} = 1.0000$, suggesting strong inferential evidence of change between the reference and current samples.

A common pattern among the highest-ranked variables is the simultaneous presence of high S_{biv} and S_{hyp} . Features ranked fourth and fifth, for example, had $S_{biv} = 1.0000$, while their hypothesis score also remained high. This indicates that the composite score is not driven solely by marginal summary-statistic changes; it also captures dependency disruption and statistically supported distributional, location, or dispersion changes. Several features in the lower part of the top ten list had relatively modest univariate scores but very high hypothesis scores, demonstrating the added value of the hypothesis component in elevating variables whose descriptive marginal deviations alone would not fully explain their instability.

Table 3. Ten highest-ranked anonymized features according to the final composite stability score. Source: Authors' own elaboration.

Rank	Feature ID	S_{uni}	S_{biv}	S_{hyp}	S_{total}	Dominant Signal
1	V_001	0.8551	1.0000	0.9990	0.9514	Bivariate
2	V_002	0.8548	1.0000	0.9990	0.9513	Bivariate
3	V_003	0.8208	0.6558	1.0000	0.8255	Hypothesis
4	V_004	0.3075	1.0000	0.9964	0.7680	Bivariate

5	V_005	0.0828	1.0000	0.9750	0.6859	Bivariate
6	V_006	0.0436	0.9485	1.0000	0.6640	Hypothesis
7	V_007	0.0329	0.9482	1.0000	0.6603	Hypothesis
8	V_008	0.0433	0.9265	1.0000	0.6566	Hypothesis
9	V_009	0.4125	0.5415	0.9998	0.6513	Hypothesis
10	V_010	0.4125	0.5415	0.9998	0.6513	Hypothesis

Table 4 reports the ten lowest-ranked variables. The lowest-ranked feature obtained a score of 0 across all components, indicating complete stability under the proposed scoring formulation. Several additional low-ranked variables had very small univariate and bivariate scores and a hypothesis score of 0, producing final scores close to zero. These features remained close to the reference state and would not be prioritized for manual review under the proposed framework.

The last three variables in Table 4 show a different pattern. Although their univariate and bivariate scores remained low, their hypothesis scores were higher, resulting in moderate but still low final scores relative to the highest-ranked variables. This illustrates the balancing effect of the composite score: statistically supported changes can increase a feature's rank, but the final score remains moderated when descriptive and dependency-related deviations are small.

Table 4. Ten lowest-ranked anonymized features according to the final composite stability score. Source: Authors' own elaboration.

Rank	Feature ID	S_{uni}	S_{biv}	S_{hyp}	S_{total}	Dominant Signal
1	V_326	0.0000	0.0000	0.0000	0.0000	Stable
2	V_325	0.0068	0.0248	0.0000	0.0105	Bivariate
3	V_324	0.0068	0.0248	0.0000	0.0105	Bivariate
4	V_323	0.0068	0.0248	0.0000	0.0105	Bivariate
5	V_322	0.0068	0.0248	0.0000	0.0105	Bivariate
6	V_321	0.0068	0.0248	0.0000	0.0105	Bivariate
7	V_320	0.0068	0.0248	0.0000	0.0105	Bivariate
8	V_319	0.0078	0.0215	0.3056	0.1116	Hypothesis
9	V_318	0.0000	0.0090	0.3496	0.1195	Hypothesis
10	V_317	0.0216	0.0316	0.3097	0.1210	Hypothesis

The contrast between Tables 3 and 4 demonstrates the diagnostic value of the component-based formulation. The highest-ranked variables are characterized by strong evidence from at least one major instability dimension, most often bivariate- or hypothesis-based. In contrast, the lowest-ranked variables remain close to the reference state across most or all components. This ranking structure supports pre-modeling feature assessment by separating variables that require further review from those that remain comparatively stable across ETL deliveries.

3.3. Hypothesis Component and *p*-Value Aggregation Results

The hypothesis-based component was examined to determine whether the highest-ranked features were supported by formal evidence of distributional, location, or dispersion change. Table 5 reports the Anderson–Darling, Mann–Whitney U, and Levene *p*-values for the ten highest-ranked variables, together with the HMP-based combined *p*-value and the transformed hypothesis score S_{hyp} . The hypothesis score is interpreted as inferential evidence of instability, where values closer to 1 indicate stronger evidence that the current feature sample differs from the reference sample.

Very small p -values are reported using threshold notation to avoid displaying numerical underflow as exact zero. In addition, Anderson–Darling p -values equal to 0.0010 indicate the lower reporting bound of the default test approximation rather than exact p -values. Therefore, the values in Table 5 are interpreted as evidence of statistically detectable feature-level differences rather than as measures of practical drift magnitude.

Table 5. Hypothesis-test evidence for the ten highest-ranked anonymized features. The main hypothesis signal corresponds to the valid test with the smallest p -value: Anderson–Darling (AD) = Distribution, Mann–Whitney U (MW) = Location, and Levene = Dispersion. Source: Authors’ own elaboration.

Rank	Feature ID	AD p -Value	MW p -Value	Levene p -Value	HMP Combined p -Value	Main Hypothesis Signal
1	V_001	0.0010	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	0.0010	Location
2	V_002	0.0010	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	0.0010	Location
3	V_003	0.0010	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	Location
4	V_004	0.0013	0.0168	0.0345	0.0036	Distribution
5	V_005	0.0093	0.0475	0.1503	0.0250	Distribution
6	V_006	0.0010	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	Dispersion
7	V_007	0.0010	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	Dispersion
8	V_008	0.0010	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	$<1 \times 10^{-15}$	Dispersion
9	V_009	0.0010	7.10×10^{-5}	0.3323	0.0002	Location
10	V_010	0.0010	7.10×10^{-5}	0.3323	0.0002	Location

The results show that all ten highest-ranked variables have S_{hyp} values above 0.975, indicating strong inferential evidence of instability. Several Mann–Whitney U and Levene p -values are reported as below 1.0×10^{-15} , while Anderson–Darling values of 0.0010 indicate the lower reporting bound of the test implementation. The dominant hypothesis signal is associated mainly with location and dispersion changes, while two variables are dominated by distributional evidence. This suggests that the highest-ranked features are not driven by a single statistical pattern; rather, different variables exhibit different forms of inferential instability.

These results also illustrate why the hypothesis component should be interpreted jointly with the descriptive Canberra-based components. Some variables in the top-ten ranking have relatively modest univariate scores but very high S_{hyp} , indicating statistically supported changes that may not be fully represented by marginal summary-statistic deviation alone. Conversely, variables with high S_{biv} and high S_{hyp} indicate both dependency-structure instability and formal evidence of distributional, location, or dispersion change. Because the reference and current ETL deliveries may partially overlap at the customer level, these p -values are interpreted as evidence of detectable differences between population snapshots rather than as paired individual-level changes.

To evaluate the choice of HMP aggregation, the HMP-based score was compared with Fisher’s, Stouffer’s, Tippett’s, and Bonferroni-type aggregation. Table 6 reports rank agreement using Spearman correlation, Kendall correlation, and top-k overlap among the highest-ranked variables. Rank-agreement metrics were computed using unrounded method scores; values in the table are rounded only for reporting.

Table 6. Rank-agreement comparison between HMP and alternative p -value aggregation methods. Source: Authors’ own elaboration.

Comparison	Spearman ρ	Kendall τ	Top-10 Overlap	Top-20 Overlap
HMP vs. Fisher	0.9884	0.9258	0.30	0.40
HMP vs. Stouffer	0.9750	0.8795	0.30	0.40

HMP vs. Tippett	0.9955	0.9627	0.40	0.60
HMP vs. Bonferroni	0.9955	0.9627	0.40	0.60

The rank correlations are high across all methods, indicating that the alternative p -value aggregation approaches produce broadly similar global feature prioritizations. HMP shows the strongest agreement with Tippett and Bonferroni-type aggregation, with Spearman correlation of 0.9955 and Kendall correlation of 0.9627. This indicates that HMP, Tippett, and Bonferroni similarly prioritize features with at least one strong hypothesis-test signal.

However, the top- k overlap values are substantially lower than the global rank correlations. For example, HMP shares only 30% of the top 10 features with Fisher and Stouffer, and 40% with Tippett and Bonferroni. This indicates that small differences in aggregation behavior can affect which variables appear in the highest-priority review list, even when the overall rankings remain highly correlated. This finding supports the comparison of multiple p -value aggregation methods rather than relying only on global rank correlation.

Figure 4 visually confirms the strong global agreement among the p -value aggregation methods. The strongest agreement is observed among HMP, Tippett, and Bonferroni, while Fisher and Stouffer also remain highly correlated with HMP. Overall, the comparison supports the use of the HMP-based score as the primary hypothesis component. HMP produces ranking behavior close to minimum- p -value-based methods while still incorporating information from all available tests through the harmonic mean formulation. The high global rank agreement but incomplete top- k overlap indicates that the choice of p -value aggregation method mainly affects the highest-priority review list; HMP was retained because it provides a balance between evidence accumulation and sensitivity to strong individual test signals. Therefore, HMP is retained in the final composite score S_{total} , while the alternative methods are used as comparative evidence for robustness and interpretation.

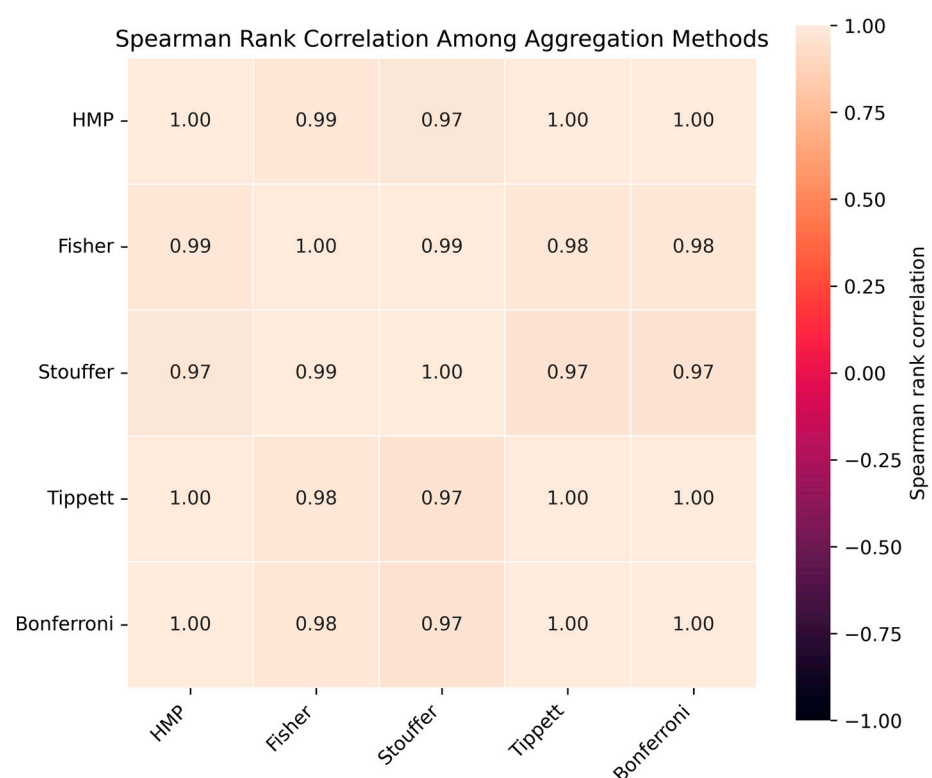


Figure 4. Spearman rank-correlation heatmap among p -value aggregation methods used to construct the hypothesis-based feature stability component. Source: Authors' own elaboration.

3.4. Composite Score Compared with Baseline and Alternative Drift Measures

The proposed composite score S_{total} was compared with the original two-component Canberra baseline score S_{base} , Jensen–Shannon divergence, and normalized Wasserstein distance. The purpose of this comparison was to evaluate whether the proposed three-component score remains consistent with the original Canberra-based framework while introducing additional ranking information through the hypothesis-based component. Because these measures operate on different mathematical scales, the comparison was performed using rank-based criteria: Spearman rank correlation, Kendall rank correlation, and top-k overlap. A supervised drift-detection accuracy was not performed because the confidential ETL deliveries do not include independent ground-truth labels indicating which features should be classified as drifted or non-drifted. Therefore, the comparison focuses on ranking agreement and top-k overlap, which are aligned with the objective of prioritizing features for pre-modeling review. Accuracy-oriented validation would require labeled drift scenarios, controlled perturbation experiments, or downstream performance labels.

The baseline score was obtained from the univariate and bivariate Canberra components defined in Section 2, while S_{total} corresponds to the final three-component formulation in Equation (12).

Jensen–Shannon divergence and normalized Wasserstein distance were computed from stored feature samples and are therefore interpreted as sample-based marginal distributional drift indicators. In contrast, S_{total} combines full EDA summary-based univariate and bivariate Canberra components with sample-based inferential evidence from the hypothesis component. The resulting rank-agreement comparison between S_{total} , S_{base} , Jensen–Shannon divergence, and normalized Wasserstein distance is summarized in Table 7.

Table 7. Rank-agreement comparison between the proposed composite score, the baseline Canberra score, and alternative distributional drift measures. Source: Authors' own elaboration.

Comparison	Spearman ρ	Kendall τ	Top-10 Overlap	Top-20 Overlap
S_{total} vs. S_{base}	0.6756	0.5744	1.00	0.80
S_{total} vs. Jensen–Shannon	0.4664	0.3311	0.30	0.40
S_{total} vs. Wasserstein	0.6195	0.4577	0.50	0.55

The comparison with S_{base} shows that the proposed score remains aligned with the original Canberra-based framework while adding new ranking information. The Spearman correlation of 0.6756 and Kendall correlation of 0.5744 indicate moderate global agreement between the two rankings. The top-k results show even stronger agreement among the highest-priority variables: the Top-10 overlap is 1.00, and the Top-20 overlap is 0.80. This means that the most critical review set is preserved after adding the hypothesis component, while the broader ranking is adjusted by inferential evidence.

The comparison with Jensen–Shannon divergence shows weaker agreement, with Spearman correlation of 0.4664, Kendall correlation of 0.3311, Top-10 overlap of 0.30, and Top-20 overlap of 0.40. This is expected because Jensen–Shannon divergence focuses on marginal distributional differences, whereas S_{total} also incorporates dependency-related deviations and hypothesis-based evidence. Therefore, lower agreement does not necessarily indicate inconsistency; rather, it shows that the proposed score captures additional dimensions of instability beyond marginal distributional divergence alone.

The normalized Wasserstein comparison shows stronger agreement with S_{total} than Jensen–Shannon divergence, with Spearman correlation of 0.6195, Kendall correlation of 0.4577, Top-10 overlap of 0.50, and Top-20 overlap of 0.55. This suggests that the composite score shares some prioritization behavior with a distributional displacement measure, particularly for features where shifts in numerical values are pronounced. However, the incomplete top-k overlap again indicates that Wasserstein distance does not fully reproduce the ranking generated by the proposed three-component framework.

To examine how the hypothesis component changed the original Canberra-based ranking, Table 8 reports the features with the largest upward and downward rank changes after adding S_{hyp} . Positive rank change indicates that the feature moved upward in priority after adding the hypothesis component, while negative rank change indicates that the feature moved downward.

Table 8. Features with the largest upward and downward rank changes after adding the hypothesis-based component. Source: Authors’ own elaboration.

Feature ID	Baseline Rank	Composite Rank	Rank Change	S_{uni}	S_{biv}	S_{hyp}	S_{base}	S_{total}	Interpretation
V_228	322	228	94	0.0000	0.0038	1.0000	0.0019	0.3346	Increased by S_{hyp}
V_229	323	229	94	0.0000	0.0034	1.0000	0.0017	0.3345	Increased by S_{hyp}
V_232	324	231	93	0.0000	0.0023	0.9999	0.0012	0.3341	Increased by S_{hyp}
V_224	314	224	90	0.0000	0.0130	1.0000	0.0065	0.3377	Increased by S_{hyp}
V_223	313	223	90	0.0000	0.0136	1.0000	0.0068	0.3379	Increased by S_{hyp}
V_255	30	255	−225	0.0000	0.5711	0.3289	0.2856	0.3000	Reduced by lower S_{hyp}
V_256	37	256	−219	0.1545	0.3938	0.3410	0.2742	0.2964	Reduced by lower S_{hyp}
V_244	27	244	−217	0.0045	0.6049	0.3433	0.3047	0.3176	Reduced by lower S_{hyp}
V_260	47	260	−213	0.0000	0.5000	0.3530	0.2500	0.2843	Reduced by lower S_{hyp}
V_252	40	252	−212	0.0084	0.5384	0.3734	0.2734	0.3067	Reduced by lower S_{hyp}

Table 8 shows that the hypothesis component redistributed feature priorities rather than uniformly increasing all scores. The first five variables had very low baseline Canberra scores, with S_{base} close to zero, but moved upward substantially because their hypothesis scores were close to 1. These cases represent statistically supported instability with limited descriptive Canberra deviation. Conversely, the last five variables had relatively high bivariate scores and therefore ranked higher under the original baseline score, but moved downward after adding S_{hyp} because their hypothesis evidence was lower relative to other features.

Overall, these results indicate that the proposed composite score preserves the main high-priority signals of the baseline Canberra framework while introducing additional ranking information through the hypothesis component. Its partial agreement with Jensen–Shannon and Wasserstein confirms that it remains related to established distributional drift measures, but it is not reducible to marginal distributional comparison. This supports the interpretation of S_{total} as a broader feature stability score that combines descriptive deviation, dependency-structure change, and inferential evidence.

3.5. Implications for Pre-Modeling Feature Selection

The proposed composite score was interpreted as a pre-modeling feature stability ranking. Unlike conventional feature selection methods, which usually evaluate predictive relevance, redundancy, or contribution to model performance, the proposed framework evaluates whether candidate features remain statistically stable across recurring ETL deliveries. This makes the score suitable as a stability-oriented screening layer before downstream model training and before model-specific feature selection is applied.

To support practical interpretation, the aligned features were grouped into descriptive stability bands according to S_{total} . These bands are used for prioritization and reporting rather than as universal decision thresholds. The four stability bands account for all 326 aligned variables. Table 9 summarizes the number and proportion of features in each stability category and provides suggested actions for model-development workflows.

The stability-band thresholds were selected as heuristic operational bands rather than universal cut-offs. They were chosen to support interpretation of the bounded composite score by separating near-zero instability ($S_{total} < 0.10$), mild instability ($0.10 \leq S_{total} < 0.25$), moderate instability ($0.25 \leq S_{total} < 0.50$), and high-priority instability ($S_{total} \geq 0.50$). In the present dataset, these thresholds also align with the observed score distribution: most features fall within the moderate interval, whereas only a limited subset reaches or exceeds 0.50 and is therefore prioritized for immediate validation. The thresholds should be interpreted as dataset- and workflow-specific reporting bands and should be recalibrated when applied to other domains, institutions, or modeling objectives.

Table 9. Feature stability bands based on the final composite score. Source: Authors' own elaboration.

Stability Band	Score Interval	Number of Features	Percentage	Suggested Action
Stable	$S_{total} < 0.10$	7	2.15	Retain; no immediate review
Mild instability	$0.10 \leq S_{total} < 0.25$	49	15.03	Monitor in next ETL cycle
Moderate instability	$0.25 \leq S_{total} < 0.50$	243	74.54	Review before modeling
High instability	$S_{total} \geq 0.50$	27	8.28	Prioritize for validation, transformation, or exclusion

Figure 5 shows the distribution of features across the four stability bands. Most variables fall into the moderate-instability category, with 243 features representing 74.54% of the aligned feature set. These variables are not necessarily candidates for immediate exclusion, but they should be reviewed before model training because they show non-negligible composite instability across the monitored dimensions. The high-instability group contains 27 features, corresponding to 8.28% of the aligned variables. These features represent the strongest candidates for validation, transformation, or possible exclusion before downstream modeling.

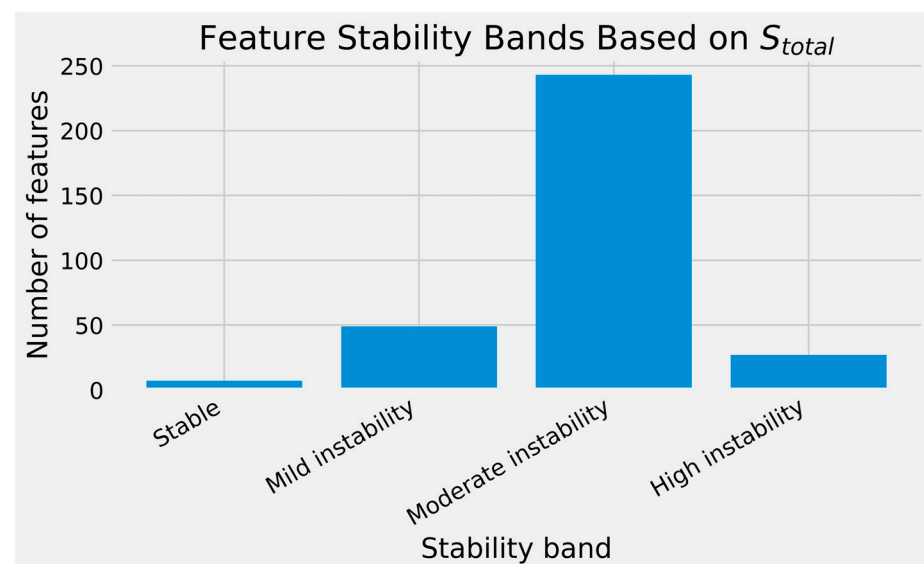


Figure 5. Feature stability bands based on S_{total} . Source: Authors' own elaboration.

Only 7 features, representing 2.15% of the aligned feature set, fall into the stable category. A further 49 features, or 15.03%, fall into the mild-instability category and may be retained while being monitored in subsequent ETL cycles. Together, the stable and mild-instability groups account for 17.18% of the variables, while the moderate- and high-instability groups account for 82.82%. This indicates that most variables exhibit some degree of instability under the proposed composite scoring framework.

These results should not be interpreted as indicating that all moderate- or high-instability variables should be removed from downstream models. A stable feature may still be irrelevant to the modeling target, and an unstable feature may still contain predictive information. Instead, the proposed ranking provides a complementary feature assessment layer that identifies variables requiring additional review before predictive feature selection is performed. For example, high-ranked variables may require investigation of upstream transformations, source-system changes, missingness patterns, population shifts, or altered dependency relationships.

In this way, the proposed framework supports feature selection workflows by introducing stability as an additional pre-modeling criterion. The score can be used to prioritize feature review, guide transformation or exclusion decisions, and reduce the risk that unstable variables are selected solely because of temporary predictive relevance in one ETL cycle.

3.6. Ablation and Weight-Sensitivity Analysis

To examine the rationale of the equal-weighted fusion strategy, an ablation and weight-sensitivity analysis was performed. The ablation analysis evaluated whether the final composite score could be reproduced by individual components or by two-component variants. This analysis was motivated by the need to assess whether the three components provide complementary ranking information or whether the final score is mainly driven by one dominant component.

The single-component variants were S_{uni} , S_{biv} , and S_{hyp} . The two-component variants were defined as $S_{uni+biv} = (S_{uni} + S_{biv}) / 2$, $S_{uni+hyp} = (S_{uni} + S_{hyp}) / 2$ and $S_{biv+hyp} = (S_{biv} + S_{hyp}) / 2$. Each variant was compared with the proposed full score S_{total} using Spearman rank correlation, Kendall rank correlation, and top-k overlap. Spearman and Kendall correlations evaluate global agreement across the full ranked feature list, while top-k overlap evaluates whether the same variables are selected for the highest-priority group.

Table 10 reports the component-ablation results. The single-component variants show only moderate agreement with the final S_{total} ranking. The univariate-only S_{uni} has a Spearman correlation of 0.5221 and a Kendall correlation of 0.3677 with S_{total} , while the bivariate-only score S_{biv} shows slightly stronger agreement, with Spearman and Kendall correlations of 0.6145 and 0.4914, respectively. These results indicate that marginal and dependency-related deviations each contribute to the final ranking, but neither component alone fully reproduces the proposed composite score.

The hypothesis-only score S_{hyp} shows moderate global rank agreement with S_{total} , with Spearman correlation of 0.5867 and Kendall correlation of 0.4359. However, its Top-10 and Top-20 overlap values are both 0.00. This result is important because it shows that the final S_{total} ranking is not determined by hypothesis-test evidence alone, despite the generally high values of S_{hyp} observed across the feature set. In other words, the equal-weighted formulation prevents statistically significant but descriptively weaker changes from automatically dominating the highest-priority review list.

The two-component variants provide stronger evidence for the complementarity of the score components. The original descriptive baseline $S_{uni+biv}$ preserves the full Top-10 review set and 80% of the Top-20 list, confirming continuity with the previous Canberra-based framework. However, its global agreement with S_{total} remains moderate, with Spearman correlation of 0.6756 and Kendall correlation of 0.5744, showing that the addition of S_{hyp} changes the broader ranking.

Table 10. Component-ablation comparison with the proposed full score S_{total} . Source: Authors' own elaboration.

Comparison with S_{total}	Spearman ρ	Kendall τ	Top-10 Overlap	Top-20 Overlap
S_{uni}	0.5221	0.3677	0.60	0.60
S_{biv}	0.6145	0.4914	0.70	0.65
S_{hyp}	0.5867	0.4359	0.00	0.00
$S_{uni+biv}$	0.6756	0.5744	1.00	0.80
$S_{uni+hyp}$	0.7448	0.5705	0.60	0.60
$S_{biv+hyp}$	0.9842	0.9282	0.70	0.80

The $S_{uni+hyp}$ variant shows higher global agreement than $S_{uni+biv}$, but its Top-10 and Top-20 overlap values remain 0.60. This indicates that excluding the bivariate component changes the highest-priority review set. The $S_{biv+hyp}$ variant shows the strongest global agreement with S_{total} , with Spearman correlation of 0.9842 and Kendall correlation of 0.9282. This suggests that dependency-related deviations and hypothesis-informed evidence are the main drivers of the broader composite ranking in this dataset. Nevertheless, its Top-10 overlap is 0.70, which means that omitting S_{uni} still changes 30% of the highest-priority feature list.

Overall, the ablation analysis supports the use of the full three-component formulation. No single component fully reproduces the final ranking, and even the strongest two-component variant does not completely preserve the Top-10 review list. Therefore, combining S_{uni} , S_{biv} , and S_{hyp} provides a more complete and diagnostically balanced feature stability ranking than relying on any individual component or two-component subset.

A separate weight-sensitivity analysis was then conducted to evaluate whether the proposed equal-weighted ranking is sensitive to plausible changes in component weights. Four alternative weighting schemes were considered: descriptive-prioritized weighting ($0.4S_{uni} + 0.4S_{biv} + 0.2S_{hyp}$), hypothesis-prioritized weighting ($0.25S_{uni} + 0.25S_{biv} + 0.50S_{hyp}$), bivariate-prioritized weighting ($0.25S_{uni} + 0.5S_{biv} + 0.25S_{hyp}$), and univariate-prioritized weighting ($0.5S_{uni} + 0.25S_{biv} + 0.25S_{hyp}$). The purpose of this analysis was not to optimize weights for a specific predictive model, but to assess whether the equal-weighted formulation provides a stable and transparent baseline for model-agnostic feature stability ranking.

Table 11 reports the weight-sensitivity results. All four alternative weighting schemes show high global rank agreement with the proposed equal-weighted S_{total} , with Spearman correlations ranging from 0.9332 to 0.9929 and Kendall correlations ranging from 0.8455 to 0.9501. This indicates that the overall feature ranking is relatively robust to reasonable changes in component weighting. The descriptive-prioritized and hypothesis-prioritized variants both preserve the complete Top-10 review list, while the hypothesis-prioritized variant also preserves the full Top-20 list. This suggests that increasing or decreasing the influence of S_{hyp} within the tested range does not disrupt the highest-priority review set.

Table 11. Weight-sensitivity comparison with the proposed equal-weighted score S_{total} . Source: Authors' own elaboration.

Alternative Weighting Scheme	Formula	Spearman ρ with S_{total}	Kendall τ with S_{total}	Top-10 Overlap	Top-20 Overlap
Descriptive-prioritized	$0.4S_{uni} + 0.4S_{biv} + 0.2S_{hyp}$	0.9394	0.8648	1.00	0.95
Hypothesis-prioritized	$0.25S_{uni} + 0.25S_{biv} + 0.50S_{hyp}$	0.9519	0.8895	1.00	1.00
Bivariate-prioritized	$0.25S_{uni} + 0.5S_{biv} + 0.25S_{hyp}$	0.9332	0.8455	0.80	0.90
Univariate-prioritized	$0.5S_{uni} + 0.25S_{biv} + 0.25S_{hyp}$	0.9929	0.9501	0.80	0.90

The bivariate-prioritized and univariate-prioritized variants show slightly lower top-k overlap than the descriptive- and hypothesis-prioritized variants, with Top-10 overlap values of 0.80 and Top-20 overlap values of 0.90. This indicates that placing stronger emphasis on a single descriptive component can change part of the highest-priority review list. However, the global rank correlations remain high, especially for the univariate-prioritized variant, which has the strongest global agreement with S_{total} among the tested weighting schemes.

Overall, the ablation and weight-sensitivity analyses provide empirical support for the proposed equal-weighted fusion strategy. The component-ablation results show that no individual score fully reproduces the final ranking and that omitting any component changes either the global ranking or the highest-priority feature list. The weight-sensitivity analysis shows that the proposed ranking remains broadly stable under reasonable alternative weighting assumptions. Therefore, equal weighting is retained as a transparent and reproducible default for pre-modeling feature stability assessment, while adaptive or task-specific weighting remains a promising direction for future work when downstream model-performance objectives or labeled drift outcomes are available.

4. Discussion

This study extended a previously proposed Canberra-based ETL drift monitoring framework by adding a hypothesis-informed feature stability component and evaluating the resulting composite score in the context of pre-modeling feature assessment. The original framework provided a descriptive ranking of variables using univariate and bivariate deviations derived from full exploratory data analysis outputs. The present extension preserves this structure but adds inferential evidence from two-sample statistical tests, aggregated through a harmonic-mean- p -value-based procedure. The aim was not to replace predictive feature selection methods, but to provide an additional stability-oriented screening layer for high-dimensional automated machine learning pipelines.

The results show that the proposed composite score S_{total} can separate different forms of feature instability. Some variables were prioritized because of strong marginal or dependency-related deviation, whereas others were elevated because of statistically supported distributional, location, or dispersion changes. This distinction is important in recurring ETL environments, where a feature may retain the same name and type while its statistical behavior changes between deliveries. In such cases, conventional schema validation may not detect the issue, and predictive feature selection may only identify the problem after model training or validation.

4.1. Main Findings and Contribution

The main finding of this study is that adding a hypothesis-informed component extends the original Canberra-based stability framework without replacing its descriptive foundation. The univariate and bivariate Canberra scores continue to provide interpretable measures of marginal and dependency-related deviation, while the newly introduced S_{hyp} component adds inferential evidence of distributional, location, or dispersion change. As a result, the final composite score S_{total} combines three complementary forms of information: descriptive marginal deviation, relational deviation, and statistically supported evidence of instability.

The score distributions showed that the univariate component was generally low across the aligned feature set, whereas the bivariate component was more dispersed. This indicates that dependency-related changes were more pronounced than simple marginal summary-statistic changes for a larger group of features. The hypothesis component produced the highest values overall, suggesting that many variables exhibited statistically detectable differences between the reference and current ETL samples. This pattern supports the interpretation that S_{hyp} captures a different aspect of feature instability from the descriptive Canberra-based components.

The ranked-feature results further showed that high final instability scores were not driven by a single source of evidence. Several variables were dominated by bivariate instability, while others were dominated by hypothesis-supported evidence. Conversely, the lowest-ranked variables showed consistently low component scores, indicating that the framework can also preserve stable features in the lower part of the ranking. This component-level structure makes the score diagnostically useful because it identifies not only which variables are unstable, but also which type of instability contributes most to their ranking.

The comparisons between the baseline score and alternative drift measures indicate that the proposed score preserves continuity with the original framework while adding new ranking information. Complete Top-10 overlap with S_{base} shows that the most critical variables remained prioritized, whereas incomplete agreement with Jensen–Shannon divergence and Wasserstein distance shows that S_{total} captures information beyond marginal distributional change.

Overall, the main contribution is the introduction of feature stability as a pre-modeling assessment criterion. The proposed framework does not replace predictive feature selection but provides a stability-oriented screening layer that can help identify variables requiring review before model-specific feature selection or training.

This interpretation is consistent with recent data-centric artificial intelligence and machine-learning data-quality studies, which emphasize that reliable model development depends on systematic evaluation, improvement, and maintenance of data quality throughout the machine learning lifecycle [10,25,37].

4.2. Interpretation of the Hypothesis-Informed Extension

The hypothesis-informed extension addresses one of the main limitations of the original Canberra-based framework: its primarily descriptive nature. The original score quantified how much a feature's summary statistics or dependency-related metrics changed between a reference dataset and current ETL delivery. However, it did not indicate whether the observed deviations were statistically supported. By adding S_{hyp} , the present framework introduces an inferential layer that helps distinguish between descriptive changes and changes supported by hypothesis-test evidence.

The three tests used in the hypothesis component capture complementary forms of instability. The Anderson–Darling test contributes evidence about changes in distributional shape and tail behavior, the Mann–Whitney U test contributes evidence about rank-based location shifts, and Levene's test contributes evidence about changes in dispersion.

This combination is useful in financial ETL data because variables may change in different ways. A feature may preserve its approximate central tendency while its spread changes, or it may show a shift in rank distribution without a large change in standard summary statistics. The hypothesis component therefore broadens the diagnostic scope of the framework.

The results from the top-ranked features show that S_{hyp} can reinforce or modify the interpretation of Canberra-based components. In several cases, variables with high bivariate instability also had high hypothesis evidence, suggesting that dependency disruption was accompanied by statistically detectable distributional, location, or dispersion changes. In other cases, variables with low descriptive Canberra deviation moved upward in the ranking because their hypothesis scores were close to 1. These cases are particularly important because they demonstrate that statistically supported instability can exist even when marginal summary deviations appear limited.

At the same time, S_{hyp} should not be interpreted as a drift magnitude. Very small p -values may occur when sample sizes are large, even when the practical size of the change is modest. This was visible in the present results, where the hypothesis component produced high values for many variables. For this reason, the hypothesis score was not used as a standalone ranking criterion. Instead, it was integrated with S_{uni} and S_{biv} , allowing statistically detectable differences to be moderated by descriptive and dependency-based evidence. This design prevents the final ranking from being driven entirely by p -values.

The use of HMP aggregation also supports the diagnostic purpose of the extension. The individual tests are applied to the same reference-current samples and therefore cannot be assumed to provide fully independent evidence. The HMP-based procedure provides a single combined evidence score while retaining sensitivity to strong test signals. The comparison with Fisher, Stouffer, Tippett, and Bonferroni-type aggregation showed high global rank agreement, but incomplete top- k overlap. This indicates that the choice of p -value aggregation method can affect the highest-priority review list, even when overall rankings are similar. HMP was retained because it provides balanced aggregation behavior: it is sensitive to small p -values while still incorporating information from all available tests.

The deterministic treatment of degenerate features also contributes to operational robustness. Some ETL variables may be constant in one delivery and variable in another, or constant with different values across deliveries. In such cases, formal hypothesis tests may not be applicable, but the change is still meaningful for feature stability monitoring. Assigning a deterministic score of 0 for identical constant values and 1 for constant-to-variable or different-constant transitions ensures that structurally important changes are not lost simply because a statistical test cannot be computed.

Overall, the hypothesis-informed extension should be viewed as an augmentation of the original Canberra framework rather than a replacement. The Canberra components provide interpretable measures of descriptive and relational deviation, while S_{hyp} adds inferential evidence. Together, the three components create a more complete feature stability ranking that is better suited for recurring ETL environments, where both magnitude of change and statistical evidence of change are relevant for pre-modeling feature review.

4.3. Comparison with Baseline and Alternative Drift Measures

The comparison with the original Canberra-based baseline score shows that the proposed extension preserves continuity with the original framework while adding new information. The complete overlap in the top ten features indicates that the most unstable variables identified by the original two-component score remained highly prioritized after the hypothesis component was added. This is important because it shows that the

extension does not disrupt the main operational signal of the baseline framework. Instead, it refines the broader ranking by incorporating inferential evidence.

The moderate global rank agreement between S_{total} and S_{base} also indicates that the hypothesis component is not redundant. If the proposed score had been almost identical to the baseline ranking, the added hypothesis layer would provide limited practical value. Conversely, if the rankings were completely different, the extension could be considered unstable or disconnected from the original scoring logic. The observed pattern is therefore desirable: the most critical features remain preserved, while the ordering of the wider feature set is adjusted by the additional evidence.

The rank-change analysis further supports this interpretation. Features with very low baseline Canberra scores but high hypothesis scores moved upward in the composite ranking. These features represent cases where descriptive summary-statistic and bivariate deviations were limited, but statistical testing indicated strong evidence of change. Conversely, features with relatively high bivariate scores but lower hypothesis scores moved downward. This demonstrates that the hypothesis component does not simply inflate all scores; rather, it redistributes priority according to the strength of inferential evidence.

The comparison with Jensen–Shannon divergence and normalized Wasserstein distance provides a second perspective on the behavior of S_{total} . These measures are commonly used to compare empirical distributions, but they primarily describe marginal distributional differences. The proposed composite score is broader because it incorporates univariate summary-statistic deviation, bivariate dependency-related deviation, and hypothesis-based evidence. For this reason, perfect agreement with Jensen–Shannon or Wasserstein distance should not be expected.

The weaker agreement with Jensen–Shannon divergence suggests that the composite score identifies forms of instability that are not captured by marginal information-theoretic divergence alone. Jensen–Shannon divergence is useful for comparing probability distributions, but it does not directly account for dependency-structure changes or the combined inferential evidence produced by distributional, location, and dispersion tests. Therefore, features can receive a high S_{total} even when their Jensen–Shannon score is not among the highest.

The stronger agreement with normalized Wasserstein distance suggests that part of the composite ranking is aligned with numerical distributional displacement. Wasserstein distance is sensitive to shifts in the location and spread of numerical values, which overlaps conceptually with some of the evidence captured by Mann–Whitney U and Levene’s tests. However, the incomplete top-k overlap shows that Wasserstein distance does not fully reproduce the proposed ranking. This is expected because Wasserstein distance does not include bivariate dependency information and does not combine multiple diagnostic dimensions.

To position the proposed framework relative to existing approaches, Table 12 compares it with major methodological directions from the literature, including classical feature selection, feature-selection stability analysis, data-validation systems, concept drift and dataset shift methods, and marginal distributional distance measures. The comparison shows that the proposed method is not intended to replace these approaches. Instead, it addresses a complementary pre-modeling task: ranking candidate features according to their stability across recurring ETL deliveries before model-specific feature selection or training is performed.

Table 12. Comparison of the proposed feature stability framework with related methodological approaches. Source: Authors' own elaboration.

Approach	Main Focus	Main Advantages	Limitations Relative to the Proposed Framework
Classical feature selection methods [1–5]	Predictive relevance, redundancy reduction, and model-specific feature usefulness	Directly support model optimization and dimensionality reduction	Usually do not evaluate whether features remain statistically stable across recurring ETL deliveries.
Feature selection stability methods [6]	Reproducibility of selected feature subsets across data perturbations or resampling	Highlight the importance of stable feature subsets for reliable modeling	Focus mainly on the stability of selected subsets rather than pre-modeling drift of individual ETL variables
Data-validation and data quality systems [7–10]	Schema checks, type validation, rule violations, and data-quality requirements	Effective for detecting structural errors, invalid values, and lifecycle data-quality issues	May not capture subtle marginal, dependency-related, or hypothesis-supported feature instability when schema checks pass
Concept drift and dataset shift methods [12–18]	Changes in data distributions, target relationships, or production data streams	Provide strong theoretical and practical foundations for monitoring changing data environments	May require labels, model outputs, stream assumptions, or full raw distributions, and may not produce interpretable variable-level stability components
Marginal Distributional distance measures [19–23]	Quantification of distributional differences using divergence, distance, or test statistics	Useful for measuring marginal distributional change and supporting drift diagnostics	Primarily focus on marginal distributions and do not jointly integrate univariate EDA deviations, bivariate dependency changes, and aggregated hypothesis-test evidence
Proposed hypothesis-informed stability framework	Pre-modeling feature stability ranking across recurring ETL deliveries	Combines descriptive, relational, and inferential evidence into an interpretable variable-level score	Does not replace predictive feature selection and requires further validation across public datasets, multiple ETL cycles, and downstream modeling tasks

Compared with classical feature selection, the proposed framework evaluates feature reliability rather than predictive usefulness. Compared with data-validation systems, it focuses on statistical and relational changes that may remain hidden after schema and type checks have passed. Compared with concept drift and dataset shift methods, it provides a variable-level ranking designed for recurring batch ETL outputs rather than a general stream-level drift decision. Finally, compared with Jensen–Shannon divergence, Wasserstein distance, and related marginal distributional measures, the proposed score integrates additional evidence from bivariate EDA deviations and hypothesis-test aggregation. These differences support the interpretation of S_{total} as a complementary stability-oriented feature assessment score rather than as a substitute for existing feature-selection, data-validation, or drift-detection methods.

Overall, the comparison indicates that S_{total} behaves as a broader feature stability measure rather than as a direct substitute for any single distributional distance. Jensen–Shannon divergence and Wasserstein distance remain useful reference measures, but they answer a narrower question: how much the marginal empirical distribution of a feature has changed. The proposed score addresses a wider operational question: whether a feature remains stable enough across ETL deliveries to be trusted as an input to downstream machine learning workflows.

4.4. Implications for Feature Selection

The proposed framework positions feature stability as a complementary criterion for feature selection in automated machine learning pipelines. Conventional feature selection methods usually focus on predictive relevance, redundancy, or model-specific contribution. These criteria are essential, but they do not necessarily indicate whether a feature remains reliable across recurring ETL deliveries. A variable may be highly predictive in one training cycle while being unstable because of changes in source systems, missingness patterns, transformations, or population composition. For this reason, stability assessment should be performed before or alongside predictive feature selection.

The results of the stability-band analysis show the practical relevance of this additional screening layer. Only a small proportion of variables fell into the stable band, while most variables were assigned to the moderate-instability category. This does not imply that most variables should be removed from modeling. Rather, it indicates that a large part of the feature space requires some degree of review before automated model training. The high-instability group represents the clearest set of candidates for immediate validation, transformation, or exclusion, whereas the moderate-instability group may require targeted review, monitoring, or domain-specific assessment.

This interpretation is especially important in high-dimensional financial ETL environments. When hundreds of variables are available, manual inspection of every feature is impractical. A ranked stability score provides a way to prioritize review effort. Features with high S_{total} can be examined first, while features with low scores can be retained with lower immediate concern from a stability perspective. This supports a more efficient feature governance process, particularly when new ETL deliveries are received periodically and model-development teams must decide quickly whether the data are suitable for training or scoring.

The proposed score also supports feature selection by reducing the risk that unstable variables are selected because of temporary or delivery-specific predictive relevance. In automated modeling workflows, unstable variables may appear useful during one training cycle but later degrade model reliability when their distribution or dependency structure changes. By flagging such variables before model-specific feature selection is performed, the framework can help reduce the likelihood that downstream algorithms select features that are statistically unreliable across deliveries.

At the same time, the proposed stability ranking should not be interpreted as a replacement for predictive feature selection. A stable feature may still have little or no relationship with the modeling target, and an unstable feature may still contain valuable predictive information. Therefore, the framework is best understood as a pre-modeling assessment layer. It identifies features that require review from a stability perspective, after which domain experts or automated model-selection procedures can decide whether to retain, transform, monitor, or exclude them.

In this way, the proposed method connects data drift monitoring with feature selection support. It shifts part of the feature-selection process upstream, before model training, by asking whether candidate variables are statistically reliable enough to be considered for modeling. This is particularly relevant for automated machine learning pipelines, where feature selection and model training may proceed with limited manual intervention. By providing a transparent and interpretable stability ranking, the framework contributes to more robust feature assessment in high-dimensional, production-oriented environments.

4.5. Limitations and Future Work

Several limitations should be considered when interpreting the results. First, the study was conducted on confidential financial ETL data, and the original feature names, raw observations, and business semantics could not be disclosed. Although anonymized

feature identifiers and aggregate score distributions allow the methodological behavior of the framework to be reported, they limit the extent to which individual feature-level findings can be externally interpreted. Although the anonymized reporting preserves feature types, score values, ranks, component contributions, and comparative evaluation outputs, the use of a single confidential financial ETL dataset limits external generalization. The present study therefore evaluates the proposed framework in the operational setting for which the baseline Canberra-based monitoring approach was originally developed, but it does not claim domain-independent validation. Future work should evaluate the method on public high-dimensional datasets or simulated recurring ETL scenarios with controlled drift mechanisms. Such evaluation would allow the framework to be tested across different domains, feature types, drift magnitudes, and data-generation settings. This limitation is common in applied financial-data studies, but future work should evaluate the method on additional datasets where more detailed feature semantics can be reported.

Second, the hypothesis component and the alternative distributional drift measures were computed from stored feature samples rather than complete raw feature distributions. This design preserves the computational efficiency and compact-output philosophy of the original EDA-based framework, but it means that S_{hyp} , Jensen–Shannon divergence, and Wasserstein distance should be interpreted as sample-based indicators. This trade-off is consistent with recent work emphasizing that unsupervised drift-monitoring methods in operational machine learning settings must balance statistical sensitivity, interpretability, and computational cost [18,38]. The Canberra-based univariate and bivariate scores remain based on full EDA summary metrics, whereas the inferential and distributional comparison components provide approximate evidence from sampled values. Sampling variability therefore remains a limitation of the hypothesis component because different samples could still produce small changes in individual p -values and in the ordering of features with similar S_{hyp} values. However, this effect was reduced in the present implementation by using a fixed random state during sample extraction. Future studies could compare sample-based and full-data implementations and evaluate repeated sampling, bootstrap confidence intervals, and ranking sensitivity to sample size, sampling strategy, and rare-level preservation.

Third, the statistical tests were applied as two-sample population-level comparisons between the reference and current ETL deliveries. In recurring financial datasets, some customers or contracts may appear in both deliveries, while others enter or leave the population. Therefore, the tests should be interpreted as indicators of feature-level instability across population snapshots rather than as paired record-level tests. If stable identifiers and sufficient overlap are available, future versions of the framework could incorporate paired or longitudinal tests to distinguish within-entity change from population-composition change.

Fourth, the hypothesis score reflects inferential evidence rather than drift magnitude. Large samples can produce very small p -values even when the practical size of the difference is modest. This was one reason for integrating S_{hyp} with the univariate and bivariate Canberra components rather than using it as a standalone score. Future work could incorporate effect-size measures alongside p -values, such as rank-biserial correlation, standardized median differences, robust variance ratios, or distributional effect sizes. This would allow the hypothesis component to express both statistical evidence and practical magnitude.

The hypothesis component was also limited to Anderson–Darling, Mann–Whitney U, and Levene tests. Although these tests capture complementary distributional, location, and dispersion changes, other statistical tests and drift detectors may produce different sensitivity profiles. Future work should examine additional methods such as Kolmogorov–Smirnov tests, Cramér–von Mises tests, energy distance, maximum mean

discrepancy, population stability index, and categorical distribution tests, especially for mixed-type, sparse, and highly imbalanced variables.

Fifth, the present study used equal weighting for S_{uni} , S_{biv} , and S_{hyp} . Equal weighting was chosen to preserve transparency and avoid model-specific tuning, but different operational settings may require different weighting schemes. For example, a pipeline focused on relational stability may give higher weight to S_{biv} , while a pipeline focused on inferential evidence may give greater weight to S_{hyp} . Future research could evaluate adaptive or data-driven weighting strategies, including weights based on historical drift behavior, model sensitivity, or expert-defined risk priorities.

Finally, the stability bands used in Section 3 are descriptive prioritization bands, not universal decision thresholds. The same score interval may have different operational meaning across institutions, data domains, or modeling objectives. Future work should therefore examine calibration across multiple ETL cycles, multiple institutions, and different modeling tasks. Longitudinal evaluation would also allow the framework to be linked with downstream model-performance degradation, enabling researchers to test whether high feature instability scores predict reduced model reliability over time.

The present study also did not directly evaluate whether excluding, transforming, or down-weighting high-instability features improves downstream machine learning performance. Therefore, the practical modeling value of the proposed score is inferred from its diagnostic behavior and feature-prioritization capability rather than from direct predictive-performance experiments. Future validation should link S_{total} with downstream model-level metrics, such as AUC, F1-score, accuracy, calibration error, prediction stability, and performance degradation across recurring ETL cycles. Such experiments would test whether features with high instability scores are associated with reduced model robustness and whether stability-aware feature filtering improves downstream performance.

Overall, future extensions should focus on multi-cycle validation, effect-size integration, paired-data testing where applicable, adaptive thresholding, and direct connection with downstream model performance. These directions would further strengthen the framework as a stability-aware feature assessment layer for automated machine learning pipelines.

5. Conclusions

This study extends the Canberra-based feature drift framework with a hypothesis-informed component for high-dimensional ETL-based machine learning pipelines. The extension adds an inferential layer to the original descriptive scoring approach, allowing feature instability to be assessed not only through univariate and bivariate deviations, but also through statistically supported evidence of distributional, location, and dispersion change.

The results indicate that the hypothesis component provides additional prioritization value without replacing the original Canberra-based structure. The highest-priority variables identified by the baseline framework remained prominent, while the broader ranking was refined by the inclusion of S_{hyp} . This shows that the proposed score can preserve continuity with the original monitoring approach while improving its diagnostic depth.

The comparison with Jensen–Shannon divergence and normalized Wasserstein distance further showed that the proposed composite score captures information beyond marginal distributional change. This is important in operational ETL environments, where instability may appear not only as distributional shift, but also as dependency disruption or statistically supported changes that are not fully reflected in descriptive summaries alone.

From a practical perspective, the framework provides a stability-oriented pre-modeling assessment layer. It does not replace predictive feature selection but supports it by identifying variables that should be reviewed, monitored, transformed, or potentially excluded before automated model training. This makes the method particularly relevant for high-dimensional financial pipelines where manual feature inspection is difficult, and recurring data deliveries may introduce subtle but important changes.

Although the framework was developed and evaluated in a financial ETL setting, the same logic may be useful in other domains where high-dimensional data are received in recurring batches and feature reliability is important. Potential applications include insurance risk modeling, healthcare data pipelines, fraud detection systems, credit monitoring, customer analytics, industrial sensor monitoring, and other production machine learning environments where variables may drift before model performance degradation becomes visible.

Future work should evaluate the framework across multiple ETL cycles, additional datasets, and downstream modeling tasks. Further development may include effect-size integration, adaptive weighting of score components, paired or longitudinal testing when stable record identifiers are available, and empirical assessment of whether feature instability scores predict model-performance degradation over time.

Author Contributions: Conceptualization, K.P. and A.E.; methodology, K.P. and A.E.; software, I.G., A.K. and K.P.; validation, K.P., I.G. and A.K.; formal analysis, K.P., I.G. and A.K.; investigation, K.P.; resources, I.G. and A.K.; data curation, I.G. and A.K.; writing—original draft preparation, K.P.; writing—review and editing, K.P., I.G. and A.E.; visualization, K.P.; supervision, A.E.; project administration, K.P.; funding acquisition, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the European Regional Development Fund within the Operational Programme “Bulgarian national recovery and resilience plan”, procedure for direct provision of grants “Establishing of a network of research higher education institutions in Bulgaria”, and under Project BG-RRP-2.004-0005 “Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia (IDEAS)”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study is unavailable due to privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
2. Destrero, A.; Mosci, S.; De Mol, C.; Verri, A.; Odone, F. Feature Selection for High-Dimensional Data. *Comput. Manag. Sci.* **2009**, *6*, 25–40. <https://doi.org/10.1007/s10287-008-0070-7>.
3. Kohavi, R.; John, G.H. Wrappers for Feature Subset Selection. *Artif. Intell.* **1997**, *97*, 273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X).
4. Liu, H.; Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*; Kluwer Academic Publishers: Boston, MA, USA, 1998. <https://doi.org/10.1007/978-1-4615-5689-3>.
5. Theng, D.; Bhoyar, K.K. Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowl. Inf. Syst.* **2024**, *66*, 1575–1637. <https://doi.org/10.1007/s10115-023-02010-5>.
6. Nogueira, S.; Sechidis, K.; Brown, G. On the Stability of Feature Selection Algorithms. *J. Mach. Learn. Res.* **2018**, *18*, 1–54.
7. Schelter, S.; Lange, D.; Schmidt, P.; Celikel, M.; Biessmann, F.; Grafberger, A. Automating Large-Scale Data Quality Verification. *Proc. VLDB Endow.* **2018**, *11*, 1781–1794.

8. Breck, E.; Polyzotis, N.; Roy, S.; Whang, S.E.; Zinkevich, M. Data Validation for Machine Learning. In Proceedings of the MLSys 2019, Stanford, CA, USA, 31 March–2 April 2019.
9. Polyzotis, N.; Roy, S.; Whang, S.E.; Zinkevich, M. Data Lifecycle Challenges in Production Machine Learning: A Survey. *SIGMOD Rec.* **2018**, *47*, 17–28.
10. Priestley, M.; O'Donnell, F.; Simperl, E. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *ACM J. Data Inf. Qual.* **2023**, *15*, 11. <https://doi.org/10.1145/3592616>.
11. Fawcett, T.; Provost, F. Activity Monitoring: Noticing Interesting Changes in Behavior. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 15–18 August 1999; pp. 53–62. <https://doi.org/10.1145/312129.312195>.
12. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A Survey on Concept Drift Adaptation. *ACM Comput. Surv.* **2014**, *46*, 44. <https://doi.org/10.1145/2523813>.
13. Webb, G.I.; Hyde, R.; Cao, H.; Nguyen, H.L.; Petitjean, F. Characterizing Concept Drift. *Data Min. Knowl. Discov.* **2016**, *30*, 964–994. <https://doi.org/10.1007/s10618-015-0448-4>.
14. Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; Zhang, G. Learning under Concept Drift: A Review. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 2346–2363. <https://doi.org/10.1109/TKDE.2018.2876857>.
15. Ben-David, S.; Blitzer, J.; Crammer, K.; Pereira, F. Analysis of Representations for Domain Adaptation. In Proceedings of the 20th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; pp. 137–144.
16. Bickel, S.; Brückner, M.; Scheffer, T. Discriminative Learning under Covariate Shift. *J. Mach. Learn. Res.* **2009**, *10*, 2137–2155. <https://doi.org/10.1145/1577069.1755858>.
17. Rabanser, S.; Günnemann, S.; Lipton, Z.C. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019. <https://doi.org/10.48550/arXiv.1810.11953>.
18. Lukats, D.; Zielinski, O.; Hahn, A.; Stahl, F.T. A benchmark and survey of fully unsupervised concept drift detectors on real-world data streams. *Int. J. Data Sci. Anal.* **2025**, *19*, 1–31. <https://doi.org/10.1007/s41060-024-00620-y>.
19. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. <https://doi.org/10.1214/aoms/1177729694>.
20. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. <https://doi.org/10.1109/18.61115>.
21. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2009. <https://doi.org/10.1007/978-3-540-71050-9>.
22. Peyré, G.; Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. <https://doi.org/10.1561/22000000073>.
23. Massey, F.J. The Kolmogorov–Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78. <https://doi.org/10.1080/01621459.1951.10500769>.
24. Piryankov, K.; Grigorova, I.; Karamfilov, A.; Efremov, A. Automated Data Monitoring Using a Canberra-Based Drift Score. *Appl. Sci.* **2026**, *16*, 2232. <https://doi.org/10.3390/app16052232>.
25. Zha, D.; Bhat, Z.P.; Lai, K.-H.; Yang, F.; Jiang, Z.; Zhong, S.; Hu, X. Data-centric Artificial Intelligence: A Survey. *ACM Comput. Surv.* **2025**, *57*, 152. <https://doi.org/10.1145/3711118>.
26. Wilson, D.J. The Harmonic Mean p -Value for Combining Dependent Tests. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 1195–1200. <https://doi.org/10.1073/pnas.1814092116>. Correction in *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 21948. <https://doi.org/10.1073/pnas.1914128116>.
27. Scholz, F.W.; Stephens, M.A. K-Sample Anderson–Darling Tests. *J. Am. Stat. Assoc.* **1987**, *82*, 918–924. <https://doi.org/10.1080/01621459.1987.10478517>.
28. Mann, H.B.; Whitney, D.R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60. <https://doi.org/10.1214/aoms/1177730491>.
29. Levene, H. Robust Tests for Equality of Variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*; Olkin, I., Ed.; Stanford University Press: Palo Alto, CA, USA, 1960; pp. 278–292.
30. Brown, M.B.; Forsythe, A.B. Robust Tests for the Equality of Variances. *J. Am. Stat. Assoc.* **1974**, *69*, 364–367. <https://doi.org/10.1080/01621459.1974.10482955>.
31. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, UK, 1925.

32. Stouffer, S.A.; Suchman, E.A.; DeVinney, L.C.; Star, S.A.; Williams, R.M., Jr. *The American Soldier: Adjustment During Army Life*; Princeton University Press: Princeton, NJ, USA, 1949.
33. Tippett, L.H.C. *The Methods of Statistics: An Introduction Mainly for Workers in the Biological Sciences*; Williams & Norgate: London, UK, 1931.
34. Bonferroni, C.E. Teoria statistica delle classi e calcolo delle probabilità. In *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*; Seeber: Florence, Italy, 1936; Volume 8, pp. 3–62.
35. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1904**, *15*, 72–101. <https://doi.org/10.2307/1412159>.
36. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. <https://doi.org/10.1093/biomet/30.1-2.81>.
37. Chen, H.; Chen, J.; Ding, J. Data evaluation and enhancement for quality improvement of machine learning. *IEEE Trans. Reliab.* **2021**, *70*, 831–847. <https://doi.org/10.1109/TR.2021.3070863>.
38. Werner, E.; Kumar, N.; Lieber, M.; Torge, S.; Gumhold, S.; Nagel, W.E. Towards Computational Performance Engineering for Unsupervised Concept Drift Detection—Complexities, Benchmarking, Performance Analysis. *arXiv* **2023**, arXiv:2304.08319. <https://doi.org/10.48550/arXiv.2304.08319>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.