A Drift-Aware and Fairness-Aware Explainable AI Framework for Major Pathogen Transitions

Fernando H. M de Oliveira[1], **Cleyton M. de O. Rodrigues**[1]

[1]Escola Politécnica – Universidade de Pernambuco (UPE)
R. Benfica, 455 – Madalena – Recife – PE – 50720-001 – Brazil

fhmo@ecomp.poli.br,cleyton.rodrigues@upe.br

Abstract. *Distinguishing SARS-CoV-2 from seasonal influenza is challenging due to overlapping symptoms and shifting disease prevalence. To address this problem, we propose a drift-aware and fairness-aware explainable AI framework for reliable clinical triage under non-stationary conditions. Using 21 low-cost triage features and Brazil's SIVEP-Gripe surveillance data (2023–2025), we evaluated XGBoost-based adaptation strategies during a major etiological inversion. Multi-task Learning with Label Shift prior correction achieved the best performance while improving demographic equity. Combined with calibration and temporal SHAP explanations, the framework provides a transparent, equitable, and cost-effective triage solution for dynamic public health settings.*

1. Introduction

The overlapping clinical presentations of SARS-CoV-2 and seasonal influenza continue to pose a significant diagnostic challenge, complicating early patient triage and straining global healthcare infrastructure [Han et al. 2025, Zhou et al. 2021, Zhu 2020]. While Reverse Transcription Polymerase Chain Reaction (RT-PCR) remains the diagnostic gold standard, its implementation for rapid, mass triage in resource-limited environments is often hindered by high costs and slow turnaround times [Yanamala et al. 2021], a limitation that persists in volatile public health systems despite the post-pandemic expansion of rapid antigen and point-of-care diagnostics. Consequently, clinicians still require affordable, automated screening tools capable of differentiating these pathogens using routinely collected triage-level variables. However, the diagnostic utility of individual signs and symptoms is inherently limited. Recent evidence indicates that common symptoms like cough can exhibit a diagnostic specificity as low as 13.4%, leading to a high volume of false positives in traditional clinical assessment, and established clinical prediction models based on symptoms typically yield Area Under the Curve (AUROC) values between 0.53 and 0.79, reflecting a discriminative ceiling when high-granularity laboratory biomarkers are absent [Webster et al. 2024].

This diagnostic challenge is severely compounded by non-stationary epidemiological landscapes. Beyond the concept drift introduced by emerging variants (e.g., Omicron) that blur clinical symptom profiles toward upper respiratory involvement [Han et al. 2025, Alemi et al. 2023], public health surveillance systems frequently experience massive label drift. Over short periods, the etiological landscape of severe acute respiratory infections (SARI) can undergo complete inversion, where the dominant pathogen shifts drastically from one season to the next. Static machine learning models, trained on historical data with fixed class priors, fail to generalize under these shifting prevalence

conditions. When the true prevalence of a target class diverges significantly from the training distribution, standard models suffer from severe miscalibration, leading to dangerous spikes in false negatives or false positives that can misallocate critical isolation resources.

To maintain reliability, there is an urgent need for drift-aware frameworks that go beyond passive monitoring. While recent studies have utilized Explainable AI (XAI) to detect feature-level drift, few have addressed the mathematical correction of label shift in tree-based architectures, nor have they evaluated how such epidemiological inversions disproportionately impact demographic subgroups. In clinical triage, a global drop in model performance often masks severe fairness disparities, where vulnerable populations experience drastically reduced sensitivity. A framework that merely detects drift without actively correcting prior shifts or auditing subgroup equity is insufficient for equitable public health deployment [Duckworth et al. 2021, Finlayson et al. 2021].

In this study, we utilize the Brazilian SIVEP-Gripe surveillance database[1] to capture a massive etiological inversion between 2023 and 2025, where COVID-19 prevalence plummeted from 78.9% to 27.3%, while influenza surged to 72.7%. Rather than relying on a single static algorithm, we propose a Drift-Aware and Fairness-Aware XGBoost Framework designed to actively mitigate label shift and audit demographic disparities. The approach evaluates advanced adaptation strategies and continuous drift monitoring combines Population Stability Index alerts with temporal SHAP analysis to track shifting symptom relevance and maintain model resilience. A fairness audit quantifies recall disparities across age, sex, and race cohorts, ensuring the tool does not codify systemic biases during viral turnover. Positioned as a low-cost, scalable support system, the framework delivers benchmark-aligned performance for real-time triage in volatile public health settings.

The study is organized as follows: Section 2 presents related works; Section 3 describes the materials and methods; Section 4 details the results; Section 5 outlines the discussion; and Section 6 provides the conclusions and future directions.

2. Related Works

Differentiating SARS-CoV-2 from influenza in acute care settings remains a persistent clinical challenge due to highly overlapping respiratory presentations. Systematic reviews and meta-analyses consistently demonstrate that hallmark symptoms such as cough, fever, and dyspnea lack pathognomonic specificity, often yielding diagnostic specificity below 15% for individual signs [Webster et al. 2024]. Consequently, clinical prediction models relying exclusively on triage-level categorical variables typically exhibit a discriminative ceiling, with AUROC values ranging between 0.53 and 0.79. While the integration of laboratory biomarkers (e.g., CRP, lymphocyte counts) and high-resolution vital signs (e.g., SpO, respiratory rate) can elevate performance, such data are frequently unavailable or delayed in resource-constrained primary triage environments [Yanamala et al. 2021]. This diagnostic ambiguity is further exacerbated by viral evolution; the emergence of Omicron lineages significantly altered the clinical phenotype of SARS-CoV-2, shifting presentations toward upper respiratory involvement and increasing overlap with seasonal influenza [Alemi et al. 2023][Han et al. 2025].

[1] https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SRAG/dicionario-de-dados-2019-a-2025.pdf

Gradient-boosting methods, particularly XGBoost [Chen and Guestrin 2016], became widely adopted in epidemiological prediction tasks due to their robustness on tabular data, computational efficiency, and native handling of missing values. Ensemble learning approaches have demonstrated strong performance for predicting severe acute respiratory infection outcomes, hospitalization risk, and pathogen classification using large-scale surveillance datasets [Viana dos S. Santana et al. 2023, de Oliveira and Rodrigues 2026, Zoabi et al. 2021]. Nevertheless, the deployment of such models in healthcare remains challenged by dataset shift, as most studies assume independent and identically distributed (i.i.d.) data and rely on random cross-validation, which often overestimates real-world performance under changing epidemiological conditions [Finlayson et al. 2021]. Rapid fluctuations in pathogen prevalence can induce severe label drift and concept drift, causing substantial degradation in predictive accuracy. Although techniques such as sample reweighting and prior-adjustment methods have been extensively studied in domain adaptation, their application to public-health surveillance and tree-based models remains limited.

Recent advances in explainable artificial intelligence (XAI) have promoted the use of SHAP to interpret clinical machine learning models [Lundberg and Lee 2017], yet most applications remain static and rarely assess temporal changes in feature importance. Similarly, the Population Stability Index (PSI), a widely used indicator of covariate shift, is seldom incorporated into clinical machine learning pipelines for continuous drift monitoring. Another critical limitation is the limited consideration of algorithmic fairness in pathogen forecasting systems. Surveillance databases such as SIVEP-Gripe often exhibit heterogeneous data quality and demographic representation [Costa et al. 2020], potentially leading to unequal predictive performance across vulnerable populations. While drift-correction methods may improve calibration and overall performance, they can also introduce unintended subgroup disparities. To date, few studies have jointly addressed label-shift correction, temporal explainability, drift monitoring, and fairness auditing within a unified framework for epidemiological surveillance.

Existing studies typically address one dimension of the problem, for instance clinical differentiation, static ML prediction, post-hoc explainability, or isolated fairness checks. None simultaneously integrate: (1) strict temporal lock-box evaluation capturing real-world prior inversion; (2) explicit label-shift correction strategies; (3) post-hoc probability calibration to restore clinical utility; (4) continuous PSI and temporal SHAP monitoring to decouple covariate from concept drift; and (5) demographic recall disparity auditing to prevent bias amplification. This study directly addresses this gap by proposing a unified, drift-aware, and fairness-audited framework designed for non-stationary public health surveillance.

3. Materials and Methods

Figure 1 illustrates the comprehensive drift-aware machine learning pipeline developed to decode clinical signatures and mitigate label shift within the Brazilian SIVEP Gripe surveillance database. This study utilized nationwide Severe Acute Respiratory Infection (SARI) surveillance data from the Brazilian SIVEP-Gripe database for patients hospitalized between January 1, 2023, and December 31, 2025. We restricted the analysis to patients with laboratory-confirmed COVID-19 (CLASSI_FIN = 5) or Influenza (CLASSI_FIN = 1), excluding cases attributed to other respiratory pathogens or unspec-

ified etiologies to isolate the specific diagnostic signals of the two primary epidemic drivers. The modeling pipeline incorporated 21 low-cost triage-level features spanning basic clinical symptoms, comorbidities, demographics, and temporal onset markers.

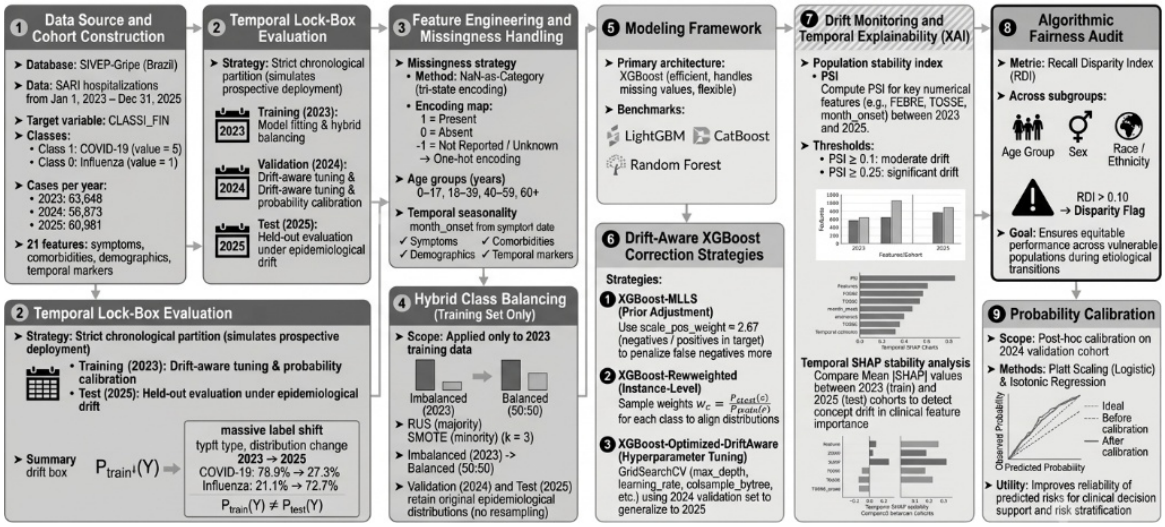


Figure 1. Overview of the proposed drift-aware machine learning pipeline for pathogen classification using the Brazilian SIVEP-Gripe surveillance database.

The proposed framework adopts a strict temporal lock-box evaluation strategy to emulate prospective clinical deployment and explicitly capture epidemiological drift. The dataset was partitioned chronologically into a training cohort (2023), a validation cohort (2024), and an independent test cohort (2025). This design avoids the unrealistic assumption of independent and identically distributed observations inherent to random cross-validation. Feature engineering was designed to preserve clinically relevant information while addressing heterogeneous missingness. Because several comorbidity variables exhibited high rates of missing data, a *NaN-as-Category* strategy was adopted, encoding symptoms and comorbidities as tri-state variables representing presence, absence, or unknown status prior to one-hot encoding. Age was grouped into four demographic categories (0–17, 18–39, 40–59, and 60+ years), and a temporal seasonality feature (`month_onset`) was extracted from symptom onset dates to capture cyclical viral transmission patterns.

To address severe class imbalance in the training set, a hybrid resampling strategy combining Random Under-Sampling (RUS) and SMOTE ($k = 3$) was applied exclusively to the 2023 training cohort, achieving a balanced (50:50) distribution of samples per class. The validation and test cohorts retained their original epidemiological distributions, ensuring an unbiased assessment of model robustness under real-world drift conditions. Although Random Forest, CatBoost, and LightGBM were evaluated as benchmark ensembles, XGBoost was selected as the primary modeling framework due to its efficient handling of missing values and flexibility for drift-aware optimization. Three complementary correction strategies were investigated: (i) a Multi-task Learning framework with Label Shift correction, using `scale_pos_weight` to adjust for changing class priors based on the target-domain ratio of negative (Influenza) to positive (COVID-19) cases, thereby penalizing false negatives in the positive class, (ii) a instance-level sample reweighting

based on the ratio between target and source class distributions, and (iii) a drift-aware hyperparameter optimization procedure that leveraged the 2024 validation cohort to improve generalization to the 2025 epidemiological inversion.

Model transparency and data-health monitoring were addressed through explainable artificial intelligence (XAI) techniques. Population Stability Index (PSI) was computed to quantify feature distribution changes between training and testing periods, while temporal SHAP analysis compared feature attributions derived from the 2023 and 2025 cohorts. This approach enabled the identification of both covariate drift and concept drift by assessing whether the clinical importance of predictive features remained stable over time. To ensure equitable performance across demographic groups, a fairness audit was conducted using the Recall Disparity Index (RDI) across age, sex, and race/ethnicity subgroups. Finally, because drift-correction mechanisms can distort probability estimates, post-hoc probability calibration was performed on the 2024 validation cohort using Platt Scaling and Isotonic Regression, improving the reliability of predicted risks for downstream clinical decision support and risk stratification.

4. Results

This section details the framework's performance during a period of major etiological inversion between 2023 and 2025. We evaluate not only global discrimination and clinical explainability but also the specific efficacy of advanced label shift correction techniques and their divergent impacts on algorithmic fairness across demographic subgroups.

4.1. Epidemiological Transition and Feature Drift (2023–2025))

Analysis of the SIVEP-Gripe study population (Table 1) reveals a profound epidemiological transition between 2023 and 2025. While the total number of laboratory-confirmed cases remained stable, the underlying demographic and etiological profiles underwent a significant shift. A pronounced demographic rejuvenation was observed, with the mean age of hospitalized patients declining from 52.3 to 44.5 years. Concurrently, the prevalence of major comorbidities, such as cardiopathy (29.0% to 23.7%) and diabetes (18.7% to 15.5%), declined in alignment with the younger patient distribution.

The most critical finding in this period is the complete etiological inversion (label drift). SARS-CoV-2, which accounted for 78.9% of the cohort in 2023, plummeted to 27.3% by 2025, while influenza-associated cases increased more than threefold, from 21.1% to 72.7%. To quantify data drift, we computed the Population Stability Index (PSI) for key clinical features between the 2023 training baseline and the 2025 test cohort. All monitored features exhibited PSI values well below the 0.1 alert threshold (e.g., Fever: 0.024, Cough: 0.030, Dyspnea: 0.009, Diabetes: 0.007). This indicates that the marginal distributions of symptoms remained stable. The model's challenge was not covariate shift, but rather severe label drift (the massive prior probability inversion) coupled with underlying concept drift (shifting feature-target relationships).

4.2. Global Performance and the Impact of Label Shift Correction

To evaluate the framework's robustness against the substantial label drift, we compared our optimized XGBoost baseline against three advanced drift-correction strategies (XGBoost-MLLS (Multi-task Label Shift correction via prior adjustment), XGBoost-Reweighted (instance-level sample reweighting), and XGBoost-Optimized-DriftAware

Table 1. Characteristics of the study population (2023–2025)

Characteristic	2023 (n = 63,648)	2024 (n = 56,873)	2025 (n = 60,981)
Age, mean (SD), years	52.3 (31.6)	48.3 (32.7)	44.5 (33.0)
Sex, n (%)			
Male	31,370 (49.3)	27,596 (48.5)	29,063 (47.7)
Female	32,273 (50.7)	29,271 (51.5)	31,913 (52.3)
Race/ethnicity, n (%)			
White	31,326 (49.2)	29,395 (51.7)	30,038 (49.3)
Black	2,252 (3.5)	1,843 (3.2)	1,941 (3.2)
Brown	19,874 (31.2)	17,341 (30.5)	21,396 (35.1)
Other/Unknown	10,196 (16.0)	8,294 (14.6)	7,606 (12.5)
Symptoms, n (%)			
Fever	34,598 (54.4)	35,824 (63.0)	40,311 (66.1)
Cough	43,116 (67.7)	42,019 (73.9)	48,702 (79.9)
Sore Throat	9,240 (14.5)	8,161 (14.3)	9,378 (15.4)
Dyspnea	35,788 (56.2)	32,455 (57.1)	36,569 (60.0)
Diarrhea	5,167 (8.1)	5,160 (9.1)	5,226 (8.6)
Vomiting	6,471 (10.2)	6,428 (11.3)	7,270 (11.9)
Loss of Smell	1,165 (1.8)	1,069 (1.9)	1,092 (1.8)
Comorbidities, n (%)			
Cardiopathy	18,443 (29.0)	15,217 (26.8)	14,430 (23.7)
Diabetes	11,909 (18.7)	9,591 (16.9)	9,463 (15.5)
Obesity	2,232 (3.5)	1,885 (3.3)	2,110 (3.5)
Etiology, n (%)			
SARI due to Influenza	13,442 (21.1)	25,694 (45.2)	44,344 (72.7)
SARI due to COVID-19	50,206 (78.9)	31,179 (54.8)	16,637 (27.3)

(hyperparameter tuning on the 2024 validation set)). Ensemble benchmarks (LightGBM, Random Forest, CatBoost) were also evaluated (Table 2). Matthews Correlation Coefficient (MCC)[Matthews 1975] and the Brier Score [Rufibach 2010] are primary indicators of classification quality and calibration.

While generic hyperparameter optimization (XGBoost-Optimized) failed to improve upon the baseline, explicit label shift correction strategies revealed distinct clinical trade-offs. XGBoost-MLLS, which adjusted the loss function to match the 2025 test priors, achieved the highest clinical sensitivity (Recall = 0.832) and F1-score (0.447). However, this came at the cost of probability calibration (Brier Score = 0.400). Post-hoc calibration via Isotonic Regression successfully reduced the MLLS Brier Score to 0.264 without altering its rank-order discrimination.

XGBoost-Reweighted achieved the best raw calibration (Brier = 0.230) and highest precision (0.361) but severely compromised sensitivity (Recall = 0.420). In a triage setting, where false negatives are critical, this reweighting strategy is unsafe. The baseline XGBoost and ensemble models maintained moderate performance, but the 2025 confusion matrix highlights a persistent safety concern. The baseline model incorrectly

Table 2. Performance Comparison of Machine Learning Models and Drift-Correction Strategies on the 2025 Test Set

Model	AUROC	AUPRC	Precision	Recall	F1	MCC	Brier
XGBoost (Baseline)	**0.606**	0.357	0.322	0.661	0.433	0.125	0.295
XGBoost-MLLS	0.604	0.354	0.305	**0.832**	**0.447**	0.124	**0.264***
XGBoost-Reweighted	0.605	0.355	**0.361**	0.420	0.388	**0.134**	**0.230**
XGBoost-Optimized	0.605	0.356	0.322	0.662	0.433	0.125	0.294
LightGBM	0.605	**0.362**	0.323	0.635	0.428	0.122	0.268
Random Forest	0.604	0.360	0.322	0.628	0.426	0.119	0.268
CatBoost	0.600	0.352	0.325	0.626	0.427	0.122	0.245

*Brier score for XGBoost-MLLS after post-hoc Isotonic Regression calibration.

classified over 23,000 true Influenza patients as having COVID-19 (Influenza specificity 47.8%). These results emphasize that while the framework maintains stability during label drift, explicit prior correction (MLLS) is required to maximize case detection.

4.3. Temporal Explainability: SHAP Stability and Concept Drift

To ensure clinical transparency and track concept drift, we utilized SHAP to decode the underlying logic of the XGBoost model and compared global feature attributions between the 2023 training cohort and the 2025 test cohort (Table 3).

Demographic variables (Age 0–17, Age 60) emerged as the most influential and stable predictors across both years. Clinically, the SHAP summary identifies cough, fever, and dyspnea as major contributors, but the temporal comparison reveals significant concept drift. The predictive importance of core respiratory symptoms (e.g., Cough decreased by 0.055, Fever by 0.070) and comorbidities significantly diminished in 2025. This temporal decay in symptom relevance confirms that the clinical presentation of the diseases evolved, likely driven by shifting viral pathophysiology. In contrast, loss of smell remained a highly stable and pathogen-specific signal (= -0.001), acting as a reliable clinical anchor even as other symptom profiles converged.

4.4. Fairness Assessment: Mitigating Disparities through Drift Correction

Global performance metrics often mask severe disparities in vulnerable subgroups. The baseline fairness audit (Table 4) revealed heterogeneous performance, with the most pronounced imbalance in model sensitivity (recall). While recall reached 0.992 in individuals aged 60 years, it dropped sharply to 0.073 among those aged 0–17 years and to 0.200 for Indigenous populations.

Our drift-correction strategies had radically divergent impacts on algorithmic equity (Table 5). The XGBoost-Reweighted model, despite its superior calibration, amplified these disparities, dropping Indigenous recall to 0.067 and pediatric recall to 0.013. By aggressively penalizing the majority class, the model erased detection capabilities for already underrepresented groups. The MLLS strategy substantially mitigated demographic disparities, increasing recall from 0.073 to 0.447 in pediatric and from 0.200 to 0.489 in Indigenous groups. The Recall Disparity Index (RDI), defined as the absolute recall difference between the highest- and lowest-performing subgroups, confirms that although

Table 3. Temporal Stability Analysis of Top Predictive Features (Mean Absolute SHAP Importance)

Feature	Train (2023)	Test (2025)	△ (Shift)
Age 0–17	1.049	1.053	+0.004
Age ≥60	0.352	0.331	-0.021
Cough	0.275	0.220	**-0.055**
Fever	0.211	0.141	**-0.070**
Cardiopathy	0.149	0.121	-0.028
Dyspnea	0.142	0.113	-0.029
Diarrhea	0.118	0.096	-0.022
Sore throat	0.110	0.067	-0.043
Diabetes	0.107	0.079	-0.028
Loss of smell	0.085	0.084	-0.001

Table 4. Algorithmic Fairness Assessment of the Baseline XGBoost Across Demographic Subgroups

Group	N	Recall	Precision	F1-score	AUROC
Race/Ethnicity					
White	30,038	0.706	0.304	0.425	0.604
Black	1,941	0.769	0.368	0.497	0.649
Brown	21,396	0.573	0.348	0.433	0.613
Indigenous	298	**0.200**	0.333	0.250	0.539
Sex					
Male	29,063	0.633	0.340	0.442	0.617
Female	31,913	0.687	0.309	0.426	0.596
Age Group					
0–17 years	21,841	**0.073**	0.375	0.122	0.603
18–39 years	4,936	0.477	0.373	0.419	0.588
40–59 years	7,728	0.760	0.306	0.437	0.596
≥60 years	26,390	0.992	0.319	0.483	0.581

Table 5. Impact of Drift-Correction Strategies on Demographic Recall Disparities (2025 Test Cohort)

Model	Age 0–17	Age ≥60	Indigenous	White	Black
XGBoost (Baseline)	0.073	0.992	0.200	0.706	0.769
XGBoost-Reweighted	0.013	0.707	0.067	0.455	0.472
XGBoost-MLLS	**0.447**	**0.999**	**0.489**	**0.858**	**0.891**

gaps persist, MLLS reduces the risk of minority groups being overlooked under label drift.

5. Discussion

This study highlights the challenges of maintaining machine learning robustness during major epidemiological transitions. The modest discrimination achieved by all models (AUROC ≈ 0.605) reflects the intrinsic overlap between COVID-19 and Influenza clinical presentations when only triage-level variables are available. Similar limitations have been reported in prior studies, where symptom-based prediction models showed restricted diagnostic performance due to the low specificity of individual respiratory symptoms [Webster et al. 2024, Alemi et al. 2023]. Consequently, the absence of physiological measurements and laboratory biomarkers imposes a natural ceiling on predictive performance.

Although AUROC remained stable across baseline and drift-corrected models, substantial differences emerged in recall, precision, calibration, and fairness, demonstrating that discrimination alone is insufficient for evaluating robustness under severe label drift. Hyperparameter optimization on the transitional 2024 cohort failed to improve performance, suggesting that conventional tuning cannot compensate for prior probability shifts. In contrast, explicit drift-correction strategies produced distinct trade-offs. The Reweighted model achieved the best probability calibration but substantially worsened subgroup fairness, whereas the MLLS approach maximized sensitivity and improved demographic equity at the cost of severe probability miscalibration. Isotonic Regression improved the MLLS Brier Score from 0.400 to 0.264, although it remained above the prevalence-only baseline (Brier 0.198), indicating residual calibration error due to the limited discriminative power of triage-only features.

Drift monitoring further revealed that covariate stability does not necessarily imply predictive stability. While PSI indicated minimal changes in symptom distributions, temporal SHAP analysis revealed substantial shifts in feature attribution, including reduced predictive importance of cough and fever. These changes may reflect true concept drift or the mathematical effects of substantial label shift and should therefore be interpreted cautiously. The findings demonstrate the value of combining PSI and SHAP to distinguish covariate from concept drift. Although MLLS improved case detection, its false positives may lead to inefficient resource allocation and delayed influenza management. Thus, it should be used as a decision-support tool rather than a substitute for laboratory diagnosis, particularly given the absence of comparisons with expert clinical judgment.

Data quality remains an important limitation. High missingness in comorbidity variables required the adoption of a NaN-as-Category strategy, which effectively handled non-random missingness without introducing imputation bias. However, the predictive value of missingness itself may become unstable if reporting practices change, creating a potential source of future model degradation. Finally, the study evaluated a single large-scale drift event involving only two respiratory pathogens within the Brazilian SIVEP-Gripe system. The lack of external validation and exclusion of other circulating respiratory viruses limit generalizability. The proposed binary classifier supports low-cost clinical triage during COVID-19 and Influenza co-circulation by enabling rapid patient stratification while laboratory results are pending.

.

6. Conclusion

This study presents a drift-aware, fairness-audited, and explainable machine learning framework specifically designed to maintain diagnostic reliability during the complex, non-stationary conditions of large-scale public health surveillance. By utilizing routinely collected triage data from the Brazilian SIVEP-Gripe database (2023–2025), our findings confirm that static models are inherently vulnerable to massive epidemiological label drift, such as the complete etiological inversion where SARS-CoV-2 prevalence plummeted from 78.9% to 27.3%. We demonstrate that standard machine learning practices, such as generic hyperparameter optimization on transitional validation data, are fundamentally insufficient to counteract these prior probability shifts.

To address this, we evaluated advanced XGBoost-based label shift correction strategies, uncovering a complex trilemma between sensitivity, calibration, and algorithmic equity. We found that explicit prior correction via Multi-task Label Shift (MLLS) adjustment maximized clinical sensitivity (Recall = 0.832) and achieved the highest F1-score (0.447). The MLLS strategy actively rescued demographic fairness, elevating recall among vulnerable Indigenous populations from 0.200 to 0.489 and pediatric cohorts from 0.073 to 0.447. In contrast, global sample reweighting achieved superior probability calibration (Brier Score = 0.230) but erased detection capabilities in these same underrepresented subgroups. To make the high-sensitivity MLLS model clinically viable, we established that a two-stage pipeline is required, applying post-hoc Isotonic Regression calibration to correct the distorted probability outputs (reducing the Brier Score from 0.400 to 0.264) without sacrificing the regained demographic equity.

Our integration of temporal Explainable AI (XAI) revealed that while covariate distributions remained stable (PSI < 0.1), profound concept drift occurred, evidenced by the temporal decay in the predictive importance of core respiratory symptoms between 2023 and 2025. This underscores the necessity of continuous SHAP-based monitoring to track evolving viral pathophysiology alongside population-level drift alerts. While the framework's global discrimination is modest (AUROC 0.605), it falls squarely within the discriminative ceiling established by international literature for secondary triage data lacking high-granularity laboratory biomarkers (AUROC 0.53–0.79). The clinical utility of the framework is accompanied by critical safety caveats. The baseline model's Influenza specificity of approximately 47.8% indicates a high volume of false positives that could compromise triage safety in isolation wards. Given its modest discrimination and focus on hospitalized SARI patients, the framework is intended as a low-cost decision-support tool for isolation-ward resource allocation, not as a standalone diagnostic tool.

Future work must transition from static bias identification to active, dynamic mitigation. This includes explore fairness-aware resampling methods to mitigate potential distortions in feature relationships, incorporate broader metrics such as Disparate Impact (DI) to assess selection rate disparities across demographic groups, integrate real-time Bayesian prior estimation to continuously update MLLS weights without requiring labeled validation sets, and implemente demographic-specific threshold optimization to guarantee equal opportunity across diverse populations.

References

Alemi, F., Guralnik, E., Vang, J., Wojtusiak, J., Peterson, R., Roess, A., and Jain, P. (2023). Guidelines for triage of covid-19 patients presenting with multisystemic symptoms. *Quality Management in Healthcare*, 32(Supplement 1):S3–S10.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Costa, L. F. L., de Mesquita Silva Montenegro, M., Rabello Neto, D. L., de Oliveira, A. T. R., Trindade, J. E. O., Adair, T., and Marinho, M. F. (2020). Estimating completeness of national and subnational death reporting in Brazil: application of record linkage methods. *Population Health Metrics*, 18(1):22.

de Oliveira, F. H. M. and Rodrigues, C. M. O. (2026). A data-driven and explainable machine learning approach to predict prolonged hospitalization in brazilian sars patients. *Revista de Engenharia e Pesquisa Aplicada*, 11(1):41–50.

Duckworth, C., Chmiel, F. P., Burns, D. K., Zlatev, Z. D., White, N. M., Daniels, T. W., Kiuber, M., and Boniface, M. J. (2021). Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Scientific Reports*, 11(1):23017.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286.

Han, Y., Guo, J., Li, X., and Zhong, Z. (2025). Differences in clinical characteristics between coronavirus disease 2019 (COVID-19) and influenza: a systematic review and meta-analysis. *NPJ Primary Care Respiratory Medicine*, 35(1):8.

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Rufibach, K. (2010). Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–939.

Viana dos S. Santana, I., Sobrinho, Á., Dias da Silva, L., and Perkusich, A. (2023). Machine learning for COVID-19 and influenza classification during coexisting outbreaks. *Applied Sciences*, 13(20):11518.

Webster, K. E., Webster, K., Parkhouse, T., Dawson, S., Jones, H. E., Jones, H., Brown, E. L., Brown, E., Hay, A. D., Hay, A., et al. (2024). Diagnostic accuracy of point-of-care tests for acute respiratory infection: a systematic review of reviews. *Health Technology Assessment*, pages 1–75.

Yanamala, N., Krishna, N. H., Hathaway, Q. A., Radhakrishnan, A., Sunkara, S., Patel, H., Farjo, P., Patel, B., and Sengupta, P. P. (2021). A vital sign-based prediction algorithm for differentiating COVID-19 versus seasonal influenza in hospitalized patients. *NPJ Digital Medicine*, 4(1):95.

Zhou, X., Wang, Z., Li, S., Liu, T., Wang, X., Xia, J., and Zhao, Y. (2021). Machine learning-based decision model to distinguish between COVID-19 and influenza: a retrospective, two-centered, diagnostic study. *Risk Management and Healthcare Policy*, pages 595–604.

Zhu, Q. (2020). On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognition Letters*, 136:71–80.

Zoabi, Y., Deri-Rozov, S., and Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digital Medicine*, 4(1):3.