CheXstray: A Real-Time Multi-Modal Monitoring Workflow for Medical Imaging AI

Jameson Merkow[1](✉), Arjun Soin[2], Jin Long[2], Joseph Paul Cohen[2],
Smitha Saligrama[1], Christopher Bridge[3,4], Xiyu Yang[3], Stephen Kaiser[1],
Steven Borg[1], Ivan Tarapov[1], and Matthew P Lungren[1,2]

[1] Microsoft Health and Life Sciences (HLS), Redmond, WA, USA
`jameson.merkow@microsoft.com`
[2] Stanford Center for Artificial Intelligence in Medicine and Imaging (AIMI), Palo Alto, CA, USA
[3] Quantitative Translational Imaging in Medicine Laboratory, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA
[4] Department of Radiology, Harvard Medical School, Boston, MA, USA

Abstract. Clinical AI applications, particularly medical imaging, are increasingly being adopted in healthcare systems worldwide. However, a crucial question remains: *what happens after the AI model is put into production?* We present our novel multi-modal model drift framework capable of tracking drift without contemporaneous ground truth using only readily available inputs, namely DICOM metadata, image appearance representation from a variational autoencoder (VAE), and model output probabilities. CheXStray was developed and tested using CheXpert, PadChest and Pediatric Pneumonia Chest X-ray datasets and we demonstrate that our framework generates a strong proxy for ground truth performance. In this work, we offer new insights into the challenges and solutions for observing deployed medical imaging AI and make three key contributions to real-time medical imaging AI monitoring: (1) proof-of-concept for medical imaging drift detection including use of VAE and domain specific statistical methods (2) a multi-modal methodology for measuring and unifying drift metrics (3) new insights into the challenges and solutions for observing deployed medical imaging AI. Our framework is released as open-source tools so that others may easily run their own workflows and build upon our work. Code available at: https://github.com/microsoft/MedImaging-ModelDriftMonitoring

Keywords: Medical imaging · Model drift · AI monitoring

1 Introduction

Recent years have seen a significant increase in the use of artificial intelligence (AI) in medical imaging, as evidenced by the rising number of academic publications and the accelerated approval of commercial AI applications for clinical

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 326–336, 2023.
https://doi.org/10.1007/978-3-031-43898-1_32

use [1,12,14,19,21]. Despite the growing availability of market-ready AI products and clinical enthusiasm to adopt these solutions [20], the translation of AI into real-world clinical practice remains limited. The reasons for this gap are multifaceted and include technical challenges, restricted IT resources, and a deficiency of clear data-driven clinical utility analyses. Efforts are underway to address these many barriers through existing and emerging solutions [4,6,13,22]. However, a major concern remains: *What happens to the AI after it is put into production?*

Monitoring the performance of AI models in production systems is crucial to ensure safety and effectiveness in healthcare, particularly for medical imaging applications. The unrealistic expectation that input data and model performance will remain static indefinitely runs counter to decades of machine learning operations research, as outlined by extensive experience in AI model deployment for other verticals [11,17]. Traditional drift detection methods require real-time feedback and lack the ability to guard against performance drift crucial for safe AI deployment in healthcare and, as such, the absence of solutions for monitoring AI model performance in medical imaging is a significant barrier to widespread adoption of AI in healthcare [7].

In healthcare, the availability of real-time ground truth data is often limited, presenting a significant challenge to accurate and timely performance monitoring. This limitation renders many existing monitoring strategies inadequate, as they require access to contemporaneous ground truth labels. Moreover, existing solutions do not tackle the distinct challenges posed by monitoring medical imaging data, including both pixel and non-pixel data, as they are primarily designed for structured tabular data. Our challenge is then to develop a systematic approach to real-time monitoring of medical imaging AI models without contemporaneous ground truth labels. This gap in the current landscape of monitoring strategies is what our method aims to fill.

In this manuscript, we present a solution that relies on only statistics of input data, deep-learning based pixel data representations, and output predictions. Our innovative approach goes beyond traditional methods and addresses this gap by not necessitating the use of up-to-date ground truth labels. Our framework is coupled with a novel multi-modal integration methodology for real-time monitoring of medical imaging AI systems for conditions which will likely have an adverse effect on performance. Through the solution proposed in this paper, we make a meaningful contribution to the medical imaging AI monitoring landscape, offering an approach specifically tailored to navigate the inherent constraints and challenges in the field.

2 Materials and Methods

2.1 Data and Deep Learning Model

We test our medical imaging AI drift workflow using the CheXpert [9] and PadChest [2] datasets. CheXpert comprises $224,316$ images of $65,240$ patients who

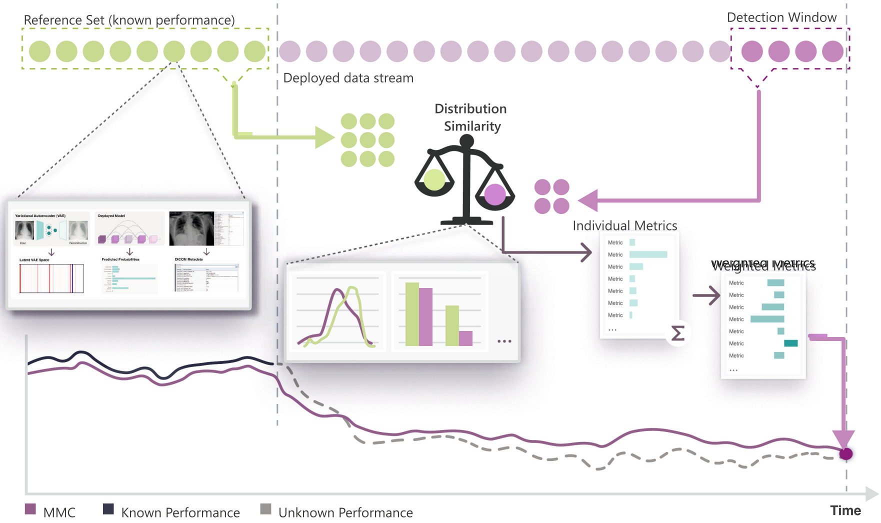


Fig. 1. Overview of our Multi-modal concordance algorithm. From each exam in datastream, we extract DICOM metadata, model predicted probabilities and a latent representation produced by a VAE then compare distributions of extracted data to a reference. We standardize and weigh these measures combining them into a value representative of the concordance to the reference.

underwent examination at Stanford University Medical Center between 2002-2017. PadChest includes 160,000 images from 67,000 patients interpreted by radiologists at San Juan Hospital from 2009-2017 with 19 differential diagnoses. Unlike other datasets, PadChest keeps the chronology of scans making it valuable for drift experiments. We categorized PadChest into three sets using examination dates: training, validation, and test. Our experimentation period spans the first year of the test set. To align with the labels in CheXpert, we consolidate relevant labels from the PadChest dataset into a set of ten unified labels. See Table 1 for details (Fig. 1).

Our approach utilizes a Densely Connected Convolutional Neural Network [8], pretrained on frontal-only CheXpert data, then we fine-tuned only the final classifier layers of the model using PadChest frontal training data. To assess the performance of our classifier over a simulated production timeframe, we employ AUROC as an evaluation metric. This approach offers a definitive indication of any potential model drift, but it necessitates real-time, domain expert-labeled ground truth labels.

2.2 Data Stream Drift and Concordance

Our approach differs from others in that we monitor for similarity or *concordance* of the datastream with respect to a reference dataset rather than highlighting differences. When the concordance metric decreases the degree to which the data

Table 1. PadChest Condensed Labels Descriptions and Distribution

	Training	Validation	Test
Start Date	2007-05-03	2013-01-01	2014-01-01
End Date	2012-12-28	2013-12-31	2014-12-31
Total Images	63,699	15,267	11,509
Pathology Counts and Descriptions			
Atelectasis	4,516	1,121	954
laminar atelectasis, fibrotic band, atelectasis, lobar atelectasis, segmental atelectasis, atelectasis basal, total atelectasis			
Cardiomegaly	5,611	1,357	935
cardiomegaly, pericardial effusion			
Consolidation	1,161	225	106
consolidation			
Edema	26	9	17
kerley lines			
Lesion	1,950	390	294
nodule, pulmonary mass, lung metastasis, multiple nodules, mass			
No Finding	21,112	4,634	3,525
normal			
Opacity	11,604	2,577	2,018
infiltrates, alveolar pattern, pneumonia, interstitial pattern, increased density, consolidation, bronchovascular markings, pulmonary edema, pulmonary fibrosis, tuberculosis sequelae, cavitation, reticular interstitial pattern, ground glass pattern, atypical pneumonia, post radiotherapy changes, reticulonodular interstitial pattern, tuberculosis, miliary opacities			
Pleural Abnormalities	6,875	1,708	1,272
costophrenic angle blunting, pleural effusion, pleural thickening, calcified pleural thickening, calcified pleural plaques, loculated pleural effusion, loculated fissural effusion, asbestosis signs, hydropneumothorax, pleural plaques			
Pleural Effusion	4,365	1,026	710
pleural effusion			
Pneumonia	3,287	584	379
pneumonia			

has drifted has increased. Our method summarizes each exam in the datasteam into an embedding consisting of DICOM metadata, pixel features and model output which is sampled into temporal detection windows in order to compare distributions of individual features to a reference set using statistical tests. Our framework, though extensible, uses two statistical tests: 1) the Kolmogorov-Smirnov (K-S) test and 2) the chi-square (χ^2) goodness of fit test. The K-S test is a non-parametric test which measures distribution shift in a real-valued sample without assuming any specific distribution [5]. The chi-square goodness-of-fit test compares observed frequencies in categorical data to expected values and calculates the likelihood they are obtained from the reference distribution

[16]. Both these tests provide a p-value and statistical similarity (distance) value. We found that statistical “distance” provided a smoother and more consistent metric which we use exclusively in our experiments, ignoring p-values.

Multi-Modal Embedding. To calculate statistics, each image must be embedded into a compressed representation suitable for our statistical tests. Our embedding is comprised of three categories: 1) DICOM metadata, 2) image appearance, and 2) model output (Fig. 2).

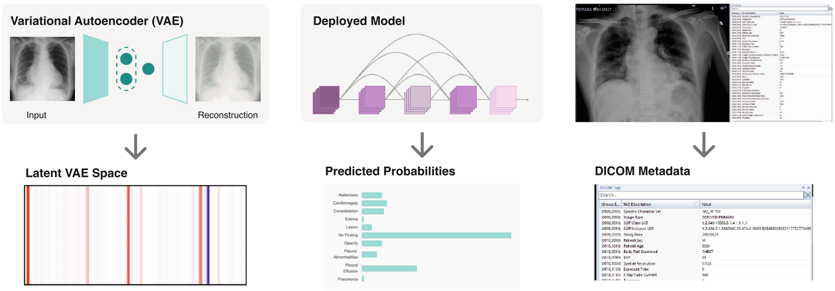


Fig. 2. Multi-Model Embedding. CheXstray utilizes data from three sources to calculate drift: 1) image appearance features (VAE), 2) model output probabilities, 3) DICOM metadata.

DICOM (Digital Imaging and Communications in Medicine), the standard for medical imaging data, includes impactful metadata like patient demographics and imaging attributes [15]. Our analysis involves key DICOM variables from the PadChest dataset, spanning patient demographics, image formation metadata, and image storage information.

AI performance in medical imaging can be affected by shifts in imaging data, due to factors like hardware changes or disease presentation variations. To address this, we employ a Variational Autoencoder (VAE)—an auto-encoder variant that models underlying parametric probability distributions of input data for fine-grained, explainable analysis [3,18,23]. Our approach uses a VAE to encode images and apply statistical tests for drift detection, representing, to our knowledge, the first VAE use case for medical imaging drift analysis. The VAE is trained on PadChest’s frontal and lateral images.

The aim of live medical data stream monitoring is to ensure consistency, detect changes impacting model performance, and identify shifts in class distribution or visual representations. We utilize soft predictions (model raw score/activation) to monitor model output and detect subtle distribution changes that hard predictions may overlook, enhancing early detection capabilities.

Metric Measurement and Unification

Our framework constructs detection windows using a sliding window approach, where the temporal parameters dictate the duration and step size of each window. Specifically, the duration defines the length of time that each window covers, while the step size determines the amount of time between the start of one window and the start of the next. We use $\hat{\psi}_i(\omega^t) = \hat{m}_i^t$ to denote individual metrics calculated at time t from a ω^t, and $\hat{m}_i^{[a,b]}$ to represent the collection of individual metric values from time a to b.

To mitigate sample size sensitivity, we employ a bootstrap method. This involves repeatedly calculating metrics on fixed-size samples drawn with replacement from the detection window. We then average these repeated measures to yield a final, more robust metric value. Formally:

$$\Theta_{\psi}(\omega, N, K) = \hat{\psi}_i(\omega) = \frac{1}{N}\sum_{0}^{N} \psi_i\left(\theta^K(\omega)\right) \tag{1}$$

where θ^K collects K samples from ω with replacement and ψ_i is the metric function calculated on the sample.

There remains three main challenges to metric unification: 1) fluctuation normalization, 2) scale standardization, and 3) metric weighting. Fluctuation normalization and scale standardization are necessary to ensure that the metrics are compatible and can be meaningfully compared and aggregated. Without these steps, comparing or combining metrics could lead to misleading results due to the variations in the scale and distribution of different metrics. We address the first two challenges by utilizing a standardization function, Γ, which normalizes each individual metric into a numerical space with consistent upper and lower bounds across all metrics. This function serves to align the metric values so that they fall within a standard range, thereby eliminating the influence of extreme values or discrepancies in the original scales of the metrics. In our experiments, we apply a simple normalization function using scale (η) and offset factors (ζ), specifically: $\Gamma(m) = \frac{m-\zeta}{\eta}$.

Metric weighting is used to reflect the relative importance or reliability of each metric in the final unified metric. The weights are determined through a separate process which takes into account factors such as the sensitivity and specificity of each metric. We then calculate our unified multi-modal concordance metric, MMC, on a detection window ω by aggregating individual metric values across L metrics using predefined weights, α_i, for each metric, as follows:

$$MMC(\omega) = \sum_{i=1}^{L} \alpha_i \cdot \Gamma_i\Big(\hat{\psi}_i(\omega)\Big) = \sum_{i=1}^{L} \frac{\alpha_i}{\eta_i}\Big(\hat{\psi}_i(\omega) - \zeta_i\Big) \tag{2}$$

where $\hat{\psi}_i(\omega)$ represents the ith metric calculated on detection window ω, Γ_i represents the standardization function, and α_i represents the weight used for the ith metric value. Each metric value is derived by a function that measures a specific property or characteristic of the detection window. For instance, one metric could measure the average intensity of the window, while another could measure the variability of intensities.

By calculating MMC on a time-indexed detection window set $\Omega^{[a,b]}$, we obtain a robust multi-modal concordance measure that can monitor drift over the given time period from a to b, denoted as $MMC^{[a,b]}$. This unified metric is advantageous as it provides a single, comprehensive measurement that takes into account multiple aspects of the data, making it easier to track and understand changes over time.

3 Experimentation

Our framework is evaluated through three simulations, inspired by clinical scenarios, each involving a datastream modification to induce noticeable drift. All experiments share settings of a 30-day detection window, one-day stride, and parameters $K = 2500$ and $N = 20$ in Eq. 1. Windows with less than 150 exams are skipped. We use the reference set for generating Ω_r and calculating η and ζ. Weights α_i are calculated by augmenting Ω_r with poor-performing samples.

Scenario 1: Performance Degradation. We investigate if performance changes are detectable by inducing degradation through hard data mining. We compile a pool of difficult exams for the AI to classify by selecting exams with low model scores but positive ground truth for their label as well as high scoring negatives. Exams are chosen based on per-label quantiles of scores, with $Q = 0.25$ indicating the lowest 25% positives and highest 25% negatives are included. These difficult exams replace all other exams in each detection window at a given point in the datastream.

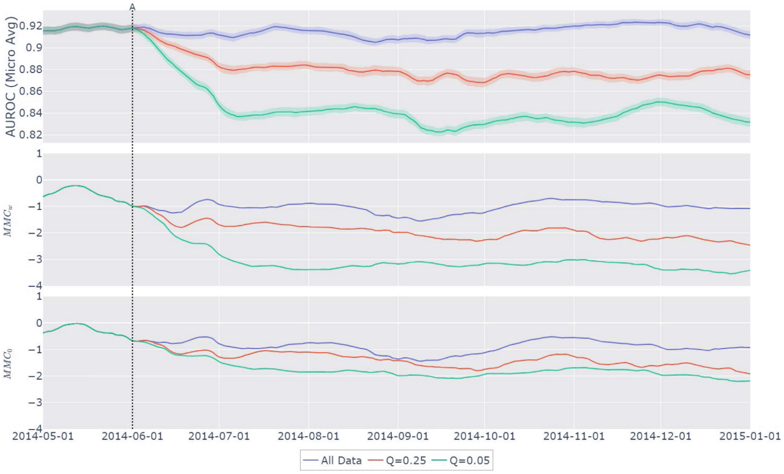
Scenario 2: Metadata Filter Failure. In this scenario, we simulate a workflow failure, resulting in processing out-of-spec data, specifically lateral images, in contrast to model training on frontal images only. The datastream is modified at two points to include and then limit to lateral images.

Scenario 3: No Metadata Available. The final experiment involves a no-metadata scenario using the Pediatric Pneumonia Chest X-ray dataset [10], comprising of 5,856 pediatric Chest X-rays. This simulates a drift scenario with a compliance boundary, relying solely on the input image. The stream is altered at two points to first include and then limit to out-of-spec data.

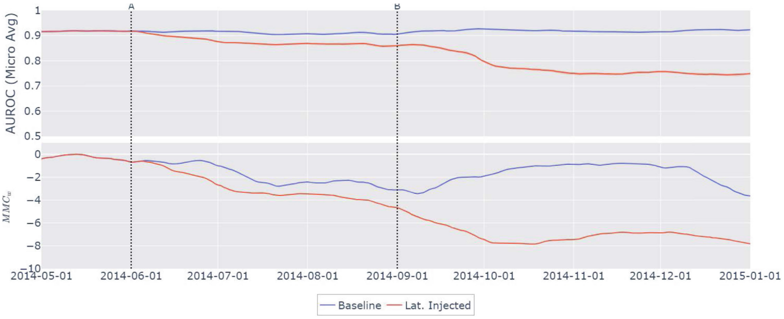
4 Results and Discussion

The performance and concordance metrics of each experiment are visualized in Fig. 3. Each sub-figure the top panel depicts performance as well as calculated MMC as calculated in each detection window. Each sub-figure has vertical lines which represent points indicating where the datastream was modified.

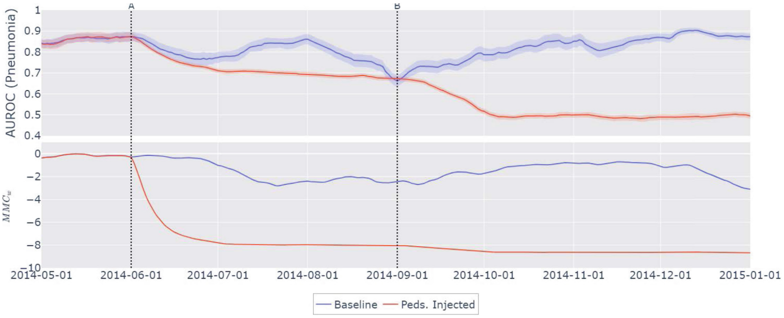
We start by discussing results in Fig. 3a from the first experiment. The top panel shows micro-averaged AUROC, the middle panel shows MMC_w, and the bottom panel shows MMC_0. As Q decreases, performance drops and the concordance metric drops as well. Both the weighted and unweighted versions of



(a) Scenario 1. Micro-average AUROC (top), MMC using weighted metric (middle), and MMC without metric weights (bottom). Baseline appears in blue. Experiments use the baseline datastream until point A where it was limited to $Q = 0.25$ (red) or $Q = 0.05$ (green) worst performing exams.



(b) Scenario 2. Micro-averaged AUROC (top) and MMC (bottom) for two data streams: the unmodified PadChest stream (blue) and a modified stream (red) with added lateral images and removed frontal images.



(c) Scenario 2. Pneumonia AUROC (top) and MMC without metadata (bottom) of unmodified PadChest stream (blue) and a modified stream (red) with added Pediatric data.

Fig. 3. Results for all three scenarios.

our metric are shown, with the weighted version providing clearer separation between performance profiles showing that our weighting methodology emphasizes relevant metrics for consistent performance proxy.

Next, Fig. 3b shows results of our second experiment. The top panel shows performance, and the bottom panel shows our metric MMC_w. Two trials are depicted, a baseline (original data stream, blue) and a second trial (red) where drift is induced. Two vertical lines denote points in time where data stream is modified: at point A, lateral images are introduced, and at point B, in-distribution (frontal) data is removed, leaving only laterals. The figure shows a correlation between performance and MMC_w; at point A, performance drops from above 0.9 to approx. 0.85 and MMC_w drops to approx. -4. At point B, performance drops to around 0.75 and MMC_w drops to and hovers around -8. This demonstrates the robustness of our method to changes in data composition detectable by metadata tags and visual appearance.

Results of our final scenario, appear in Fig. 3c. In this experiment, we measure Pneumonia AUROC, as the pediatric data includes only pneumonia labels. We observe a drop in performance and concordance at both points where we modify the data stream, show that our approach remains robust without metadata and can still detect drift. We also notice a larger drop in concordance compared to performance, indicating that concordance may be more sensitive to data stream changes, which could be desirable for detecting this type of drift when the AI model is not cleared for use on pediatric patients.

We demonstrate model monitoring for a medical imaging with CheXStray can achieve real-time drift metrics in the absence of contemporaneous ground truth in a chest X-ray model use case to inform potential change in model performance. This work will inform further development of automated medical imaging AI monitoring tools to ensure ongoing safety and quality in production to enable safe and effective AI adoption in medical practice. The important contributions include the use of VAE in reconstructing medical images for the purpose of detecting input data changes in the absence of ground truth labels, data-driven unsupervised drift detection statistical metrics that correlate with supervised drift detection approaches and ground truth performance, and open source code and datasets to optimize validation and reproducibility for the broader community.

Acknowledgments. This work was was supported in part by the Stanford Center for Artificial Intelligence in Medicine and Imaging (AIMI) and Microsoft Health and Life Sciences.

References

1. Benjamens, S., Dhunnoo, P., Meskó, B.: The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit. Med. **3**(1), 1–8 (2020)
2. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: PadChest: a large chest x-ray image dataset with multi-label annotated reports. Med. Image Anal. **66**, 101797 (2020). https://doi.org/10.1016/j.media.2020.101797