Perfect, this will be your **speaker script PDF**: one page per slide, same headers, aimed at AML students and your teachers. You can copy-paste each "page" into Word/LaTeX/Google Docs, keep one page per section, and export as PDF.

(I'm following the slide titles from your Distribution_Shift_Presentation_v18_last.pptx file.)

Page 1 – Distribution Shift in Medical Imaging

Header: Distribution Shift in Medical Imaging

*"Good morning everyone. My name is Moataz Mansour, and today I'll present our CAS AML project on **distribution shift in medical imaging**, done together with Sharad George.*

*The central problem we address is the following: we typically train deep learning models on one carefully curated dataset, for example NIH ChestX-ray14, and then later we want to deploy them in a different hospital, on a different patient population, or with a different acquisition protocol. In that case, the data the model sees in production is no longer drawn from the same distribution as the training data. This phenomenon is what we call **distribution shift**.*

In medical imaging, this is not just a theoretical concern. A model that looks excellent on its internal test set can silently degrade once we move to a new site, and this can directly affect clinical decisions and patient safety.

*In this project, we explore whether we can use an **unsupervised signal** – specifically the reconstruction error of a convolutional autoencoder trained on NIH data – to detect such distribution shifts early, before labels are even available in the new domain. We then relate this unsupervised signal to the actual performance of a supervised classifier, a DenseNet-121, on several external datasets.*

The key idea is to treat the autoencoder as a sort of "distribution sensor" and ask: when this sensor says 'these images look different', does the classifier indeed perform worse on them?"

Page 2 – Why Distribution Shift Matters

Header: Why Distribution Shift Matters

"Let me first motivate why distribution shift is such a critical issue in applied machine learning, especially in medicine.

*When we train a model, we implicitly assume that the training data and the deployment data are **independent and identically distributed** – the classic i.i.d. assumption. In reality, this rarely holds. New hospitals may use different scanners, different image post-processing, or serve a different patient population. Even something as simple as a new firmware version on an X-ray machine can change the intensity distribution of the images.*

In medical imaging, these changes can have serious consequences. A model that was calibrated on adult US patients may not perform the same on a pediatric population in China, or on images from a different institution such as Stanford. We might see a drop in sensitivity for important pathologies, or a shift in thresholds where the model becomes over-confident or under-confident.

*The second reason this matters is that **labels are expensive and delayed**. Radiologist annotations are not readily available when you move to a new site. So, we need tools that can **flag a potential problem without relying on labels**. That is what motivates the use of autoencoder reconstruction error and latent-space analysis in this project: we want an early warning signal for distribution shift – ideally one we can compute from the raw images alone."*

Page 3 – Research Question & Approach

Header: Research Question & Approach

"Based on this motivation, we framed our work around three main research questions.

*First: **Can we detect distribution shift between chest X-ray datasets in an unsupervised way, using only image data?** More concretely, if we train an autoencoder purely on NIH data, will its reconstruction error clearly distinguish NIH images from images coming from other datasets like a Chinese pediatric cohort or CheXpert from Stanford?*

*Second: **How does this unsupervised signal relate to supervised performance?** If the autoencoder says 'this dataset is far from NIH', do we actually observe a drop in classifier performance – and can we quantify that relationship?*

*Third: **Which factors seem to dominate the shift?** Is it mostly the change in pathology prevalence, the demographic differences, or the institutional/technical setup such as scanners and pipelines?*

*To answer this, our approach is organized into phases in the notebook:

1. A **Phase 1** baseline analysis using simple pixel statistics and Jensen–Shannon divergence.

2. **Phase 2**, where we train convolutional autoencoders on NIH data and understand their behavior on NIH normals vs abnormals.
3. **Phase 3**, where we evaluate reconstruction error and latent-space structure on external datasets.
4. **Phase 4**, where we introduce a DenseNet-121 classifier trained on NIH and examine its performance on the same external data.
5. **Phase 5**, where we correlate the unsupervised shift measures with the supervised performance metrics.*

The presentation follows this structure closely.”

Page 4 – Datasets Used

Header: Datasets Used

“The first ingredient in this study is the choice of datasets, which we selected to span different institutions and populations.

*Our **reference domain** is **NIH ChestX-ray14**, a large dataset of adult chest radiographs collected in the United States. It contains frontal chest X-rays with multiple pathologies labeled. We down-select and preprocess these images into a binary problem: ‘No Finding’ versus ‘Abnormal / Pneumonia’. All our models are trained on NIH only.*

*The second dataset is a **pediatric chest X-ray dataset from China**, containing children with pneumonia and healthy controls. Compared with NIH, this dataset differs in two main ways: the patients are children rather than adults, and the images come from a different institution, including different equipment and acquisition protocols.*

*The third dataset is **CheXpert** from Stanford, another large chest X-ray dataset. Here, the patient population is still mainly adult, but the **labeling process, scanners and image post-processing** differ from NIH, and the class balance is very skewed toward abnormal findings.*

By combining these three datasets – NIH, Chinese pediatric, and CheXpert – we create scenarios where we can tease apart the impact of demographics, pathology prevalence, and institutional differences on distribution shift.”

Page 5 – Label Distributions

Header: Label Distributions

*“Beyond just the images, the **label distributions** are very different across these datasets, and this has a strong impact on evaluation.*

*In **NIH**, after our preprocessing and binary mapping to ‘No Finding’ vs ‘Abnormal’, the dataset is **relatively balanced**, roughly around half normals and half abnormals. This is a reasonable setting for training a binary classifier and for interpreting plain accuracy.*

*In the **Chinese pediatric dataset**, the situation is different: a much larger fraction of cases are labeled as pneumonia, so the normal class is the minority. This already tells us that a model trained on NIH may see a different mixture of pathologies and severities.*

*In **CheXpert**, the imbalance is extreme: only a very small fraction of images are truly ‘normal’, and the vast majority are abnormal. A naive classifier that simply predicts ‘abnormal’ for everything would achieve a deceptively high **standard accuracy** on CheXpert, even though it completely fails to recognize normal cases correctly.*

*These differences in label distribution are exactly why we later move away from plain accuracy and use **balanced accuracy**, and why we interpret AUC with caution. They also motivate Phase 5, where we look at how these class balances interact with reconstruction error and performance.”*

Page 6 – Phase 1 – Baseline Distribution Shift

Header: Phase 1 – Baseline Distribution Shift

*“Phase 1 is our sanity check: before using any learned model, we ask whether we can already see signs of distribution shift using **simple pixel statistics**.*

*In the notebook, we compute intensity histograms or pixel distributions for each dataset and then measure the divergence between them using **Jensen–Shannon (JS) divergence**. JS divergence is a symmetric and smoothed version of KL divergence that gives us a number between 0 and 1, where 0 means identical distributions and larger values indicate greater dissimilarity.*

We perform this analysis both on the full datasets and on subsets restricted to ‘normal’ images, to separate pathology effects from pure acquisition/processing effects.

*The purpose of this phase is not to provide a perfect domain-shift measure, but to set a **baseline**: if even the pixel intensity distribution already shows a clear separation between NIH, Pediatric, and CheXpert, then it is entirely plausible that higher-level features learned by a neural network will also feel this shift.”*

Page 7 – Phase 1 – Key Findings

Header: Phase 1 – Key Findings

“The key findings from Phase 1 can be summarized as follows.

*First, the **JS divergence** between NIH, Pediatric and CheXpert is clearly **non-zero** and of substantial magnitude. This means that, at a very low level, the pixel distributions differ – which is exactly what we would expect if the scanners, protocols, and post-processing pipelines are not identical.*

*Second, when we repeat the analysis using **only normal images**, the divergences remain clearly visible. This is important: it tells us that the shift is **not solely driven by pathology**. Even if we remove disease as a factor and look only at presumably healthy lungs, the datasets still look different in terms of intensity distributions.*

*This gives us two takeaways:
– Distribution shift is already detectable without learning anything, just from pixels.
– Non-pathology factors such as **institutional and technical differences** play a major role.

These observations motivate the next phase, where we move from marginal pixel distributions to learned representations via a convolutional autoencoder trained on NIH.”

Page 8 – Phase 2 – Convolutional Autoencoder

Header: Phase 2 – Convolutional Autoencoder

*“In Phase 2 we introduce our main unsupervised tool: a **convolutional autoencoder** trained on NIH chest X-rays.*

*An autoencoder consists of two parts: an **encoder**, which maps the input image to a low-dimensional latent representation, and a **decoder**, which reconstructs the image from this representation. We use a convolutional architecture, so both encoder and decoder are composed of convolutional and transposed-convolutional layers.*

*We resize each image to 224×224 and treat it as a single-channel input. The encoder progressively downsamples the image through convolution and pooling, and finally compresses it into a **256-dimensional latent vector**. The decoder then mirrors this process and reconstructs back to 1×224×224.*

*We train this model on NIH data only, minimizing the **mean squared error** between the input and the reconstruction. During training, the autoencoder learns to capture the dominant structures in NIH chest X-rays: lungs, heart, ribs, and typical noise patterns.*

*Importantly, we train two variants: one on **all NIH images** (NIH_Full AE) and one on **only NIH normal images** (NIH_Normal AE). The second variant is intended to learn* *specifically what "healthy NIH chest" looks like, making deviations more interpretable later on."*

Page 9 – Phase 2 – Convolutional Autoencoder

Header: Phase 2 – Convolutional Autoencoder

"This second autoencoder slide is where I briefly comment on the ***design choices and practical training aspects*** *that are documented in the notebook.*

We chose a moderately deep architecture with around ***tens of millions of parameters****, which is expressive enough to reconstruct chest X-rays but still trainable on the available hardware. We use standard initialization, batch-based training with Adam, and a learning rate that we tune to get stable convergence without overfitting.*

Because the autoencoder is fully convolutional, it preserves spatial locality. This is important in medical images, where structures like lung fields and heart silhouettes have consistent locations. It also means that the model can share filters across the whole image, keeping parameter count manageable compared to a fully connected model.

During training, we monitor ***training and validation reconstruction loss****. The notebook shows that both losses decrease smoothly and plateau, indicating that the autoencoder is indeed learning meaningful structure and not just memorizing the training set.*

Once we are confident that the model has converged on NIH, we treat the encoder and decoder as ***frozen*** *and move to the next step: analyzing reconstruction error and latent codes for various datasets."*

Page 10 – Phase 2 – Training Results

Header: Phase 2 – Training Results

"On this slide we look at ***quantitative training curves*** *for the autoencoder.*

The notebook shows typical loss curves: starting from a relatively high reconstruction error, both training and validation losses drop and then flatten. We do not observe a large gap between them, which suggests that the autoencoder generalizes reasonably well to unseen NIH images. This is important, because we want the reconstruction error to reflect how 'typical' an image is for the NIH domain, not just how similar it is to specific training samples.

We also observe that the ***NIH_Normal AE*** *achieves particularly low error on NIH normals, as expected. When we feed it NIH* ***abnormal*** *images, the reconstruction error* *increases slightly. In the notebook, this difference is around* ***+6–7%*** *relative to normals. We interpret this as the* ***baseline pathology effect*** *within the NIH distribution: abnormal lungs are a bit harder for a normal-only autoencoder to reconstruct, but they are still within the same hospital and acquisition setup.*

These baseline numbers are crucial because they become our yardstick later: if cross-dataset errors are 100–200% higher, we can confidently say that institutional and demographic shifts dominate over pathology alone."

Page 11 – Phase 2 – Training Results

Header: Phase 2 – Training Results

"Here, I focus on the ***qualitative reconstructions*** *shown in the notebook.*

We can visually compare input images and their reconstructions for NIH normals and abnormals. On NIH normals, the autoencoder reproduces the global lung structure, heart contour, and rib patterns quite faithfully. Fine-grained textures might be slightly smoothed, but overall the reconstruction is very close to the original.

For NIH abnormals, especially those with strong opacities or consolidations, we see that the autoencoder sometimes 'softens' or partially smooths out the pathological regions. This is typical: the model has learned a strong prior on what a "typical" chest X-ray looks like and tends to reconstruct more average-looking lungs. However, it still integrates many of the abnormal patterns, leading to a modest increase in reconstruction error.

The key point is that the ***autoencoder has clearly internalized the NIH distribution****. Within that domain, both normals and abnormals are still 'familiar' images. This will be in stark contrast with what happens when we feed images from Pediatric or CheXpert, which the autoencoder has never seen and which come from different imaging pipelines."*

Page 12 – Phase 2 – Training Results

Header: Phase 2 – Training Results

"This slide gives us room to emphasize the interpretation of those training results in the context of distribution shift.

First, the small difference in error between NIH normals and abnormals in the NIH_Normal AE – on the order of a few percent – shows that within a single institution, pathology alone causes only a ***moderate increase*** *in reconstruction difficulty. In other* *words, abnormal lungs are unusual, but not extremely out-of-distribution with respect to the model's prior.*

*Second, the stability of the loss curves suggests that the autoencoder is **not overfitting** to idiosyncratic artifacts. This is important because we want the reconstruction error to be a robust measure of how typical an image is, not just how similar it is to a small training set.*

*Third, these results justify using reconstruction error as a **continuous 'distance' measure** from the NIH training distribution. We can now meaningfully ask: if an image from another dataset produces an error that is, say, 100% higher than NIH baseline, is that a sign that the image lies far outside the NIH domain? Phase 3 is precisely about quantifying that."*

Page 13 – Phase 2 – Training Results

Header: Phase 2 – Training Results

"On this last training-results slide, I connect Phase 2 back to the bigger picture of the project.

*You can think of the trained autoencoder as a **compact model of the NIH distribution**. The encoder defines a latent manifold of NIH-like images, and the decoder knows how to move from that latent space back to image space. Reconstruction error then becomes our proxy for how far a new image is from this learned manifold.*

*So Phase 2 has two main outcomes:

1. It gives us a well-trained model whose reconstruction error we can now apply to any chest X-ray, regardless of its origin.
2. It quantifies, within NIH, how much reconstruction error changes between normal and abnormal cases – our internal reference for 'how big is big' when we later see external shifts.*

*With this foundation in place, we can now proceed to Phase 3, where we stop training and simply **feed images from Pediatric and CheXpert into the frozen autoencoder** to see how unusual they look from the point of view of the NIH model."*

Page 14 – Phase 2 – Key Findings

Header: Phase 2 –Key Findings

"To close Phase 2, let me summarize the key findings.

*First, the convolutional autoencoder trained on NIH converges well and reconstructs NIH images with **low error**, confirming that it has captured the essential structure of the NIH domain.*

*Second, when we restrict training to NIH normals, the resulting autoencoder treats NIH abnormals as slightly more difficult, with around **+6–7%** higher reconstruction error. This is our internal **pathology baseline**.*

Third, the combination of quantitative loss curves and qualitative reconstructions gives us confidence that the model is both expressive and stable, making its reconstruction error a meaningful signal.

*These findings justify using reconstruction error as a candidate **unsupervised domain-shift measure**. The next question, tackled in Phase 3, is: how does this error behave when we go beyond NIH and step into truly new datasets?”*

Page 15 – Phase 3 – Reconstruction Error Across Datasets

Header: Phase 3 – Reconstruction Error Across Datasets

*“In Phase 3, we keep the autoencoder weights **fixed** and systematically change only the **input dataset**. This is where we test our core idea: if reconstruction error reflects distance from the NIH distribution, it should increase substantially when we feed images from Pediatric and CheXpert.*

The evaluation protocol is as follows. We take the trained NIH_Full AE and NIH_Normal AE and compute reconstruction error for three types of inputs:
– Images from the NIH dataset (our reference).
– Images from the Chinese pediatric dataset.
– Images from CheXpert.

*We compute metrics such as the **mean reconstruction error**, but also look at distributions – box plots, histograms – to see how the entire error distribution shifts. For NIH_Normal AE, we pay special attention to **normal-only subsets** of each dataset, so we can isolate how much of the shift is due to pathology versus institutional and demographic differences.*

This phase produces the first strong numerical evidence that reconstruction error is indeed sensitive to distribution shift.”

Page 16 – Phase 3 – Quantitative Results

Header: Phase 3 – Quantitative Results

“The first quantitative result is already striking.

*When we apply the **NIH_Full AE** and compare **mean reconstruction error** across datasets, we see that:*
*– NIH images, as expected, have the **lowest error**.*
*– Pediatric images show an error that is roughly on the order of **+100% higher** than NIH baseline.*
*– CheXpert images show an even larger increase, roughly **+200% or more** relative to NIH.*

*These increases are **orders of magnitude larger** than the +6–7% we saw between NIH normals and NIH abnormals in Phase 2. This tells us that, from the autoencoder’s point of view, ‘this is a different hospital’ or ‘this is a different population’ is a much bigger shock than ‘this patient has pneumonia’.

Statistical tests such as the Mann–Whitney U test confirm that these differences in error distributions are highly significant. We are not just looking at noise; the entire reconstruction-error distribution shifts upward.

So already at this point we can say: reconstruction error is strongly correlated with being ‘out of NIH domain’.”

Page 17 – Phase 3 – Quantitative Results

Header: Phase 3 – Quantitative Results

*“The next step is to repeat this analysis specifically in the **normal-only setting** using the NIH_Normal AE.*

We restrict inputs to images labeled as normal in each dataset and compute reconstruction error using the normal-only autoencoder. The idea is to remove obvious pathology as a confounder and ask: if we look at presumably healthy lungs from each dataset, do they still look very different to the NIH model?

The answer is yes. Normal Pediatric images produce significantly higher reconstruction error than NIH normals, and normal CheXpert images are again the most difficult for the NIH_Normal AE to reconstruct. The relative increases remain very large compared with the NIH abnormal vs normal baseline.

*This is a crucial finding: **even healthy lungs from other datasets are far from the NIH notion of ‘normal’**. This strongly supports the hypothesis that institutional and technical differences dominate the shift, with demographics as an additional factor in the pediatric case.*

In other words, the autoencoder is a very sensitive detector of these cross-domain differences, even when we remove pathology from the picture.”

Page 18 – Phase 3 – Quantitative Results

Header: Phase 3 – Quantitative Results

*“Finally, we look beyond means and analyze the **full error distributions** and effect sizes.*

Boxplots and histograms in the notebook show that the entire distribution of reconstruction error for Pediatric and CheXpert is shifted upward and often broadened compared with NIH. There is very little overlap in the bulk of the distributions, and statistical tests produce extremely small p-values.

*We also compute **relative effect sizes**, for example the ratio of median error in Pediatric vs NIH and CheXpert vs NIH. These ratios, again, are far larger than the internal pathology effect within NIH. This reinforces our interpretation that what the autoencoder perceives as ‘unusual’ is primarily the domain change, not the disease itself.*

*At this point, reconstruction error has proven itself as a strong unsupervised indicator of distribution shift. The natural next question is: how does this relate to what really matters in practice – the performance of a **diagnostic classifier**?”*

Page 19 – Phase 3 – Interpretation

Header: Phase 3 – Interpretation

“Before moving on to classifiers, let me summarize the interpretation of Phase 3.

The autoencoder, trained solely on NIH data, defines a notion of what ‘typical’ NIH chest X-rays look like, encoded in both its latent space and its decoder’s ability to reconstruct. When we feed it Pediatric or CheXpert images, the model struggles much more: reconstruction error increases by more than 100% for Pediatric and more than 200% for CheXpert in some settings.

*Crucially, even normal images from these datasets are perceived as highly atypical. This means the shift is largely due to **institutional/technical and demographic factors**, not just pathology.*

From a domain adaptation perspective, this suggests that simply retraining thresholds or calibrating probabilities may not be enough. The underlying representation itself is misaligned with the new domain.

*These findings set the stage for two things:

1. Using reconstruction error as a **real-time domain shift monitor** at deployment.
2. Investigating, in Phases 4 and 5, whether higher reconstruction error is actually predictive of worse classifier performance on those datasets.*"

Page 20 – Phase 3c – Latent Space Visualization (t-SNE)

Header: Phase 3c – Latent Space Visualization (t-SNE)

*"So far we have treated reconstruction error as a single scalar per image. In Phase 3c, we look inside the model and visualize the structure of the **latent space** using **t-SNE**.*

The encoder of the NIH_Normal AE maps each image to a 256-dimensional latent vector. This space captures the compressed representation of chest X-rays according to the NIH model. To make this interpretable, we use t-SNE to embed these high-dimensional vectors into 2D, while approximately preserving local neighborhood relationships.

*On the t-SNE plot, each point is an image, colored by dataset: for example, blue for NIH, orange for Pediatric, and green for CheXpert. What we observe is that **the points cluster by dataset**. NIH images occupy one region of the map, pediatric images another, and CheXpert images a third.*

This means that even though the autoencoder was only trained on NIH, its latent space naturally separates images from other datasets as separate clusters. It confirms that the model's internal representation sees these datasets as distinct populations, consistent with the reconstruction error story."

Page 21 – t-SNE and Classifier Performance

Header: t-SNE and Classifier Performance

"Now, how does this latent-space structure connect to classifier performance?

*Intuitively, the NIH cluster in the t-SNE plot represents **in-distribution data**: this is where the classifier, trained on NIH, has seen numerous examples and learned decision boundaries that make sense. The Pediatric and CheXpert clusters represent regions of the latent space where the model has **never** seen data during training.*

*When we deploy a classifier trained purely on NIH in such regions, it is effectively **extrapolating**. Extrapolation in high-dimensional space is risky: the model may produce confident predictions, but there is no guarantee that its learned decision boundary is still valid. This is precisely what we observe when we later measure performance: balanced accuracy drops on Pediatric and even more on CheXpert.*

So the t-SNE plot is not just a pretty picture. It gives us a geometric intuition: the further a dataset's cluster is from the NIH cluster in latent space, the more we expect classifier performance to degrade. In other words, latent-space distance and reconstruction error are both proxies for how safe it is to trust a model trained on NIH."

Page 22 – t-SNE and Classifier Performance

Header: t-SNE and Classifier Performance

*"On this slide we can zoom further into the **structure within** the t-SNE clusters.*

Within the NIH cluster, we often see a gradient that roughly corresponds to pathology severity: very clean normals cluster together, while more complex abnormals lie along certain directions. However, once we move to the Pediatric or CheXpert clusters, we see that the model organizes those images differently, reflecting their own internal structure and possibly the different mix of pathologies and acquisition settings.

*If we overlay reconstruction error on the t-SNE plot, for example by coloring points by error magnitude, we see that points near the NIH cluster tend to have low error, while those far away – especially many CheXpert points – have high error. This reinforces the interpretation of reconstruction error as a **radial distance** from the NIH manifold in latent space.*

Thus, the t-SNE visualization provides strong qualitative support for our later quantitative correlation analysis: regions of latent space where data are far from NIH are exactly the regions where we expect metrics like balanced accuracy and specificity to suffer."

Page 23 – t-SNE and Classifier Performance

Header: t-SNE and Classifier Performance

*"To connect this even more directly to the classifier, we can imagine overlaying **classifier errors** on the t-SNE map.*

*If we mark misclassified points in the latent space, we typically see that most misclassifications occur **outside** the dense core of the NIH cluster, and more frequently in the Pediatric and CheXpert regions. This is consistent with the idea that the classifier is well tuned for in-distribution NIH data but less reliable in other parts of latent space.*

This leads to a conceptual picture of deployment:
– When a new image falls near the NIH cluster in latent space and has low reconstruction error, we can be relatively confident in the classifier prediction.
– When a new image falls in a region dominated by Pediatric or CheXpert points and

exhibits high reconstruction error, we should treat the prediction with caution and consider options like human review, threshold adjustment, or retraining.

In practice, we would not run t-SNE in real time, but reconstruction error is readily available and correlates with these regions, making it a practical proxy for how 'far' a case is from the training domain."

Page 24 – t-SNE and Classifier Performance

Header: t-SNE and Classifier Performance

*"This slide gives room to emphasize how these visual insights feed into our **model-selection and model-acceptance** thinking.*

Suppose a hospital wants to evaluate a third-party chest X-ray model. They can use an autoencoder trained on their own local data to compute latent embeddings and reconstruction errors for validation cases. If the external model's training data cluster far away from the hospital's cluster, or if reconstruction errors are systematically high, this is a red flag that the model may not generalize well to the local population.

In our study, we use NIH as the reference and treat Pediatric and CheXpert as external domains. The t-SNE plots and reconstruction errors tell us that these external domains are indeed far away. Later, in Phase 4 and Phase 5, we see the consequences: the NIH-trained classifier performs clearly worse on these domains than on NIH itself.

*So, the t-SNE analysis is not just academic; it informs **practical decision-making** about where a model can safely be deployed and where additional adaptation or caution is needed."*

Page 25 – t-SNE and Classifier Performance

Header: t-SNE and Classifier Performance

"To close the t-SNE section, let me summarize the main lessons for AML students and practitioners.

*First, latent-space visualization is a powerful tool to **debug and understand** models, especially in multi-dataset settings. It provides an intuitive geometric view of how the model organizes data and where distribution shifts occur.*

*Second, t-SNE should not be over-interpreted quantitatively – distances are distorted – but **cluster structure and neighborhood relationships** are still informative. In our case, the fact that NIH, Pediatric and CheXpert form distinct clusters strongly supports the conclusion that the model sees them as different domains.*

Third, combining t-SNE with reconstruction error and classifier errors gives a multi-faceted picture: where are the data in latent space, how far are they from the training manifold, and how often does the classifier fail there? This combination is more informative than any single metric alone.

With this in mind, we move on to Phase 4, where we introduce the actual classifier architecture and look at its performance across these domains."

Page 26 – Phase 4 – DenseNet-121 Classifier

Header: Phase 4 – DenseNet-121 Classifier

"In Phase 4 we focus on the supervised classifier and ask: given what we've learned about distribution shift, how well does a model trained on NIH perform on each dataset?

We use a ***DenseNet-121*** *architecture, pre-trained on ImageNet, as our feature extractor. We adapt it to grayscale chest X-rays by either duplicating the single channel or adjusting the first convolution to accept one channel, and we resize images to 224×224 to match the expected input size.*

We freeze the early dense blocks, which capture low-level features, and fine-tune the later blocks and classification head on NIH. The output layer is a single sigmoid unit predicting the probability of pneumonia/abnormal vs no finding. During training, we use class-balanced sampling or class weights to handle the mild imbalance in NIH, apply standard data augmentation (flips, small rotations, slight zoom), and monitor validation performance.

It is very important to stress that ***this classifier does not use the autoencoder's latent features****. The autoencoder is a completely separate module, used only for the unsupervised shift analysis. DenseNet sees raw images, and its performance is evaluated independently.*

With the classifier trained, we evaluate it on NIH, Pediatric and CheXpert and study metrics like AUC and balanced accuracy."

Page 27 – Phase 4 – Classifier Performance

Header: Phase 4 – Classifier Performance

"On this slide we show the main performance results of the DenseNet-121 classifier across the three datasets.

*As expected, the model performs **best on NIH**, the domain it was trained on. Both AUC and balanced accuracy are relatively high, indicating good discrimination between normal and abnormal cases and a reasonable trade-off between sensitivity and specificity.*

*When we evaluate on the **Pediatric** dataset, we typically see a drop in performance: sensitivity or specificity – and thus balanced accuracy – decline. This reflects the fact that children's lungs and acquisition protocols differ from what the model has seen before.*

*On **CheXpert**, if we look only at **standard accuracy or even AUC**, the model can appear to perform surprisingly well, partly because the dataset is heavily skewed toward abnormal cases. However, when we look at **balanced accuracy** and especially specificity on the minority normal class, we see that performance is much worse. The model tends to over-call abnormal, which is rewarded by accuracy but punished by balanced metrics.*

These results connect directly back to the autoencoder story: the datasets that produced the highest reconstruction errors – Pediatric and especially CheXpert – are precisely the ones where the classifier's balanced performance is weakest."

Page 28 – Evaluation Metrics – AUC and Accuracy

Header: Evaluation Metrics – AUC and Accuracy

*"To interpret these results correctly, we need to be very clear about our **evaluation metrics**.*

*First, **AUC**, the area under the ROC curve, measures how well the model ranks positive cases above negative ones across all possible thresholds. It is threshold-independent and robust to some forms of imbalance, but it does not tell us directly about performance at a specific operating point, nor does it reflect the actual error rates in deployment.*

*Second, **plain accuracy** – the fraction of correct predictions at a chosen threshold – can be highly misleading in imbalanced settings. For example, in CheXpert, where almost all cases are abnormal, a classifier that simply predicts 'abnormal' for everything will achieve very high accuracy, yet is clinically useless at detecting truly normal cases.*

*In our results, we observed exactly this effect: AUC and accuracy on CheXpert may look superficially good, but this hides a serious deficiency on the minority class. This is why we introduce **balanced accuracy** as a fairer metric, especially when comparing across datasets with very different label prevalences."*

Page 29 – Slide 29

Header: Slide 29 (ROC/curve illustration – narration)

“On this slide you see the ROC curves or related visualizations that illustrate how the classifier behaves on each dataset.

For NIH, the ROC curve is pleasantly far from the diagonal, indicating solid discriminative ability. For Pediatric and CheXpert, we may still see reasonably high AUC values, but the shape of the curve and the underlying confusion matrices reveal that the model pays a price in either sensitivity or specificity, especially at clinically relevant thresholds.

In particular, in an imbalanced dataset like CheXpert, the ROC curve can still look good even when the classifier is biased toward the majority class. That’s why, in addition to AUC, we look at metrics that give ***equal weight to both classes****, such as balanced accuracy, and at per-class performance measures.*

The overall message of this slide is that AUC alone is not sufficient to judge model robustness under distribution shift. We need metrics that reflect the ***trade-offs*** *between false positives and false negatives in each domain.”*

Page 30 – Balanced Accuracy – Fair Evaluation

Header: Balanced Accuracy – Fair Evaluation

“Balanced accuracy is our main tool to correct for class imbalance and to fairly compare performance across datasets.

*Formally, for a binary classifier, we define:
– **Sensitivity (True Positive Rate)**: the proportion of actual positives that are correctly identified.
– **Specificity (True Negative Rate)**: the proportion of actual negatives that are correctly identified.

Balanced accuracy is simply the ***average of sensitivity and specificity****:*
[\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}.]

This means that each class contributes equally to the metric, regardless of how frequent it is. A classifier that predicts everything as abnormal on CheXpert will have very high sensitivity but near-zero specificity, resulting in a balanced accuracy close to 0.5 – which clearly reflects its poor performance on normals.

In our project, balanced accuracy reveals that the NIH-trained model’s performance deteriorates significantly on Pediatric and CheXpert, even when AUC remains *acceptable. This metric therefore gives a more honest picture of how distribution shift harms real-world performance."*

Page 31 – Slide 31

Header: Slide 31 (Confusion / threshold illustration – narration)

"Here we can interpret confusion matrices or threshold-specific metrics for the three datasets.

On NIH, at a reasonable operating threshold, the classifier maintains both high sensitivity and high specificity, leading to strong balanced accuracy. The confusion matrix shows relatively few false negatives and false positives.

On the Pediatric dataset, we typically see an increase in either false negatives or false positives, depending on how the threshold is tuned. Clinically, this could correspond to missing pneumonia cases or over-calling disease in healthy children, both of which are problematic.

On CheXpert, especially at thresholds that maintain good sensitivity, specificity plummets on the minority normal class. This again is not so visible in AUC or raw accuracy, but it becomes clear in balanced accuracy and in the confusion matrix.

This slide reinforces the importance of picking ***appropriate metrics and thresholds****, and it visually aligns with our reconstruction-error and t-SNE findings: the more out-of-distribution the data, the more skewed the confusion matrices become."*

Page 32 – Phase 5 – Correlation Analysis

Header: Phase 5 – Correlation Analysis

"In Phase 5 we explicitly study the relationship between the ***unsupervised shift signal*** *and the* ***supervised performance metrics****.*

The core idea is simple: if reconstruction error is a good proxy for 'distance' from the training distribution, then cases or groups of cases with higher error should tend to have ***worse classifier performance****. To test this, we compute reconstruction error for images, aggregate them into bins or per-dataset summaries, and correlate these with metrics like balanced accuracy, sensitivity, and specificity.*

For example, we can compare the average reconstruction error of NIH, Pediatric and CheXpert with their corresponding balanced accuracies. We observe a ***negative correlation****: NIH has the lowest error and the highest balanced accuracy, while*

CheXpert has the highest error and the lowest balanced accuracy, with Pediatric in between.

This is an encouraging result: it suggests that an unsupervised signal, available even before labels, carries information about how well the classifier is likely to perform."

Page 33 – Phase 5 – Correlation Analysis

Header: Phase 5 – Correlation Analysis

"We can also look at more fine-grained analyses within a dataset or across subsets of data.

*One option, as illustrated in the notebook, is to stratify images by reconstruction-error quantiles and then compute performance metrics within each quantile. Typically, we see that quantiles with **low reconstruction error** have better classifier performance, while quantiles with **high error** have worse performance. This holds not only across datasets but also within them, especially when we mix samples from different domains.*

*Another perspective is to compute rank-correlations or simple regression lines between average error and metrics like balanced accuracy at the dataset level. Although we only have a few datasets here, the trend is consistent: **higher error → lower balanced accuracy**.*

Phase 5 therefore provides quantitative evidence that reconstruction error is not just reflecting some arbitrary difference; it is systematically associated with how reliable the classifier is on those images. This is exactly what we want from a domain-shift monitor."

Page 34 – Limitations, Future Work & Takeaways

Header: Limitations, Future Work & Takeaways

*"Of course, our study has several **limitations** that are important to acknowledge, especially in an academic setting.*

*First, we restrict ourselves to **chest X-ray data** and to a specific selection of datasets. The conclusions may not generalize directly to other modalities such as CT, MRI, or ultrasound, or to other clinical tasks.*

Second, we focus on a single autoencoder architecture and a single classifier family, DenseNet-121. Different architectures or training regimes might produce different sensitivities to distribution shift.

Third, our correlation analysis is based on a relatively small number of domains and summaries. To build a fully predictive model of performance from reconstruction error, we would need more datasets and more systematic experimentation.

Finally, our work is purely retrospective; we do not integrate clinical feedback or prospectively deploy the system in a real hospital setting. So, while the signals are promising, their practical impact still needs to be validated in partnership with clinicians."

Page 35 – Limitations, Future Work & Takeaways

Header: Limitations, Future Work & Takeaways

"In terms of ***future work****, there are several natural extensions.*

One direction is to explore ***alternative unsupervised shift detectors****: for example, other generative models like variational autoencoders or diffusion models, or contrastive/self-supervised encoders trained on NIH. We could compare how well their reconstruction error, likelihood, or feature distances correlate with performance.*

Another direction is to build a more detailed ***performance-prediction model****: given reconstruction-error distributions and some meta-information about a new dataset, can we forecast, with uncertainty, the expected balanced accuracy or the need for retraining?*

We could also investigate ***per-pathology analyses****: does reconstruction error predict performance drop equally well for different disease categories, or are some pathologies more robust than others under distribution shift?*

Finally, integrating this into a ***real MLOps pipeline*** *– with dashboards, automatic alerts, and retraining triggers – would turn this from a research project into a concrete safety mechanism for deployed models."*

Page 36 – Limitations, Future Work & Takeaways

Header: Limitations, Future Work & Takeaways

"Let me summarize the key takeaways for AML students and teachers.

First, ***distribution shift is inevitable*** *in real deployments, and simple i.i.d. assumptions rarely hold in medicine. Ignoring this can lead to over-optimistic evaluations and unsafe models.*

Second, ***unsupervised tools*** *like convolutional autoencoders provide valuable signals about how different new data are from the training distribution. Reconstruction error* *and latent-space structure, visualized with techniques like t-SNE, can reveal cross-institution and demographic differences even without labels.*

Third, in our experiments, reconstruction error correlates with ***classifier performance degradation*** *across datasets: the more the autoencoder struggles, the worse the DenseNet-121 performs in terms of balanced accuracy.*

Fourth, ***metric choice matters****: AUC and accuracy can look good even when a model is failing on the minority class, especially under extreme imbalance. Balanced accuracy and per-class metrics are essential for fair evaluation under distribution shift.*

Overall, the project shows that combining unsupervised shift detection with careful supervised evaluation can make medical AI systems more transparent and safer to deploy."

Page 37 – Real-World Applications

Header: Real-World Applications

"Finally, let me end with some concrete real-world applications of this approach.

First, an autoencoder trained on a hospital's historical data can be deployed alongside any classifier as a ***'watchdog' model****. It continuously computes reconstruction error for incoming images; if the error distribution drifts beyond a predefined threshold, the system raises an alert. This can trigger actions like human review, temporary model suspension, or targeted retraining.*

Second, during ***model procurement****, hospitals can evaluate off-the-shelf models on a small sample of local images, using an autoencoder to assess how far those images are from the vendor's training distribution. Models that are extremely sensitive to domain differences can be rejected or flagged for additional adaptation.*

Third, in a broader MLOps context, reconstruction error and latent-space statistics can be integrated into ***monitoring dashboards*** *that track data drift over time, helping teams distinguish between benign changes and dangerous shifts that require intervention.*

In summary, our project is not just about a specific autoencoder and a specific DenseNet; it illustrates a general pattern: use unsupervised representation learning to monitor where your model is operating in input space, and use that information to make more informed, safer decisions about when and where to trust its predictions.

Thank you for your attention."

If you like, I can next:

- Shorten or expand specific pages to match your speaking time, or
- Add explicit time estimates per slide (e.g., “≈30–45 seconds”) so you can pace a 12–15 minute talk.