Influence-Based Sample Valuation for Chest X-Ray Classification

Danial Ressl

Abstract

Large medical imaging datasets often contain substantial noise, imbalance, and redundancy, raising questions about how individual samples contribute to model learning. This project applies influence-based data valuation to the NIH ChestXray14 dataset to examine how specific images shape the behavior of a deep chest X-ray classifier. A DenseNet-121 model is trained using focal loss and label smoothing to mitigate class imbalance and noisy supervision. Influence scores are computed using TracIn, which approximates training-sample impact through gradient alignment across checkpoints (Pruthi et al., 2020). Using these scores, I conduct removal and selection experiments to assess dataset redundancy and compressibility. The results indicate that a significant fraction of the most influential samples may be removed with minimal performance loss, and that a comparatively small subset of high-influence samples is sufficient to recover much of the model's accuracy. These findings align with data-centric and human-centered perspectives in machine learning, underscoring the importance of dataset quality and interpretability in clinical applications (Shankar et al., 2023).

Visual Abstract
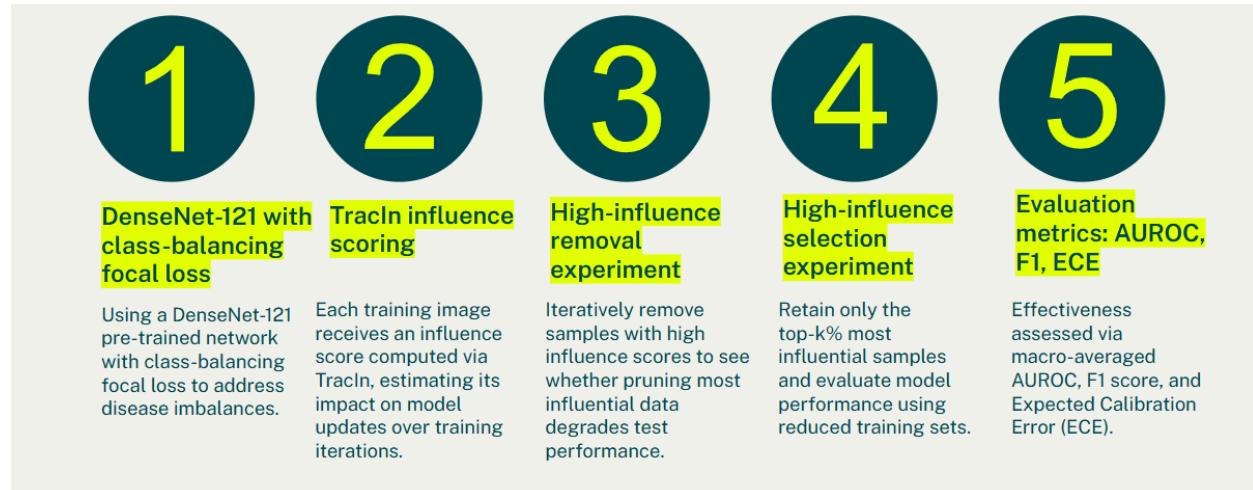
Figure 1. *Visual abstract summarizing the project workflow, including dataset preprocessing, DenseNet-121 training with focal loss, TracIn influence estimation, and influence-based removal and selection experiments.*

1 Introduction

Developing robust machine learning systems for healthcare requires close attention to the quality, provenance, and structure of training data. This challenge is particularly pressing in the Canadian healthcare system; radiologist shortages and diagnostic backlogs are just some of the motivators towards exploration of automated decision-support tools. Clinical imaging datasets however, frequently contain labeling inconsistencies and substantial class imbalance, both of which can degrade model reliability as well as complicate evaluation.

The NIH ChestXray14 dataset embodies these issues. Labels in NIH14 were produced using rule-based NLP applied to radiology reports (Wang et al., 2017), rather than through expert annotation. Various audits have shown that many labels in the dataset are noisy, unreliable, or even clinically ambiguous, especially for classifications such as infiltration, pneumonia, and consolidation (Oakden-Rayner, 2017). These characteristics make NIH14 a particularly important dataset for examining how training samples influence model learning and generalization.

Recent work in data-centric AI suggests that improving dataset quality often yields larger gains than refining model architectures. The DataPerf framework (Shankar et al., 2023) formalizes this viewpoint by emphasizing dataset evaluation and curation as primary levers for improving system performance. Complementary efforts in human-centered ML frame datasets as sociotechnical artifacts whose construction processes materially shape model behavior and downstream reliability (Shneiderman, 2020).

Influence-based methods provide a principled mechanism for interrogating these dataset issues. Classical influence functions (Koh & Liang, 2017) quantify how removing a training point would affect loss, but such methods are computationally expensive making them infeasible for deep networks. TracIn avoids this limitation by estimating influence through gradient dot products accumulated over checkpoints (Pruthi et al., 2020). This makes influence estimation tractable for large deep-learning pipelines.

This project investigates how individual samples influence learning dynamics in NIH14. A DenseNet-121 model is trained using focal loss and label smoothing, after which TracIn influence scores are computed for the entire training set. Removal and selection experiments are conducted to assess redundancy and dataset compressibility. Through these analyses, the project demonstrates how influence-based techniques can expose weaknesses in noisy clinical datasets, while supporting data-centric approaches to model improvement.

2 Methods

2.1 Dataset

The NIH ChestXray14 dataset contains 112,120 frontal chest radiographs labeled for fourteen thoracic pathologies (Wang et al., 2017). Labels were assigned using an automated keyword-based NLP pipeline, which introduces substantial noise due to ambiguous phrasing, radiological uncertainty, and report-level context not visible in images. Prior critique has shown that several labels are poorly defined or inconsistently applied, and that clinically irrelevant linguistic patterns appear as labels (Oakden-Rayner, 2017). The dataset is also highly imbalanced, with rare classes e.g., hernia appearing in less than one percent of images. This combination of noise, imbalance, and redundancy makes NIH14 well-suited for this influence-based analysis.

2.2 Model Architecture

A DenseNet-121 architecture pretrained on ImageNet is used as the baseline classifier. DenseNet models are commonly adopted in medical imaging tasks due to their strong representational efficiency and stable optimization properties. The final classification layer outputs fourteen logits corresponding to the disease labels.

2.3 Training Procedure

Training is performed using focal loss, which down-weights easy examples and emphasizes hard or ambiguous ones. This choice is appropriate for NIH14, where mislabeled or uncertain examples are common. Label smoothing (0.05) is applied to reduce overconfidence and mitigate calibration errors. Optimization uses AdamW with cosine annealing. Data augmentation includes random resized cropping, flipping, and light color jitter. The model is trained for five epochs, which is sufficient to achieve stable learning for influence scoring while avoiding overfitting to noisy labels.

2.4 Influence Estimation with TracIn

Influence estimation uses TracIn, a method that approximates the effect of training samples by summing gradient dot products across multiple checkpoints (Pruthi et al., 2020). This approach is inspired by influence functions (Koh & Liang, 2017), but avoids intractable second-order computations. For each training image, its gradient at each checkpoint is compared with validation gradients, and the accumulated score reflects whether the sample contributed positively or negatively to reducing validation loss.

Under focal loss, gradients for difficult or mislabeled examples are amplified, while gradients for easy examples are suppressed. This results in a tightly compressed influence distribution centered near zero, with both positive and negative values. Negative influence indicates that a

training point consistently pushes the model weights in a direction that increases validation loss, often indicating mislabeled or atypical samples.
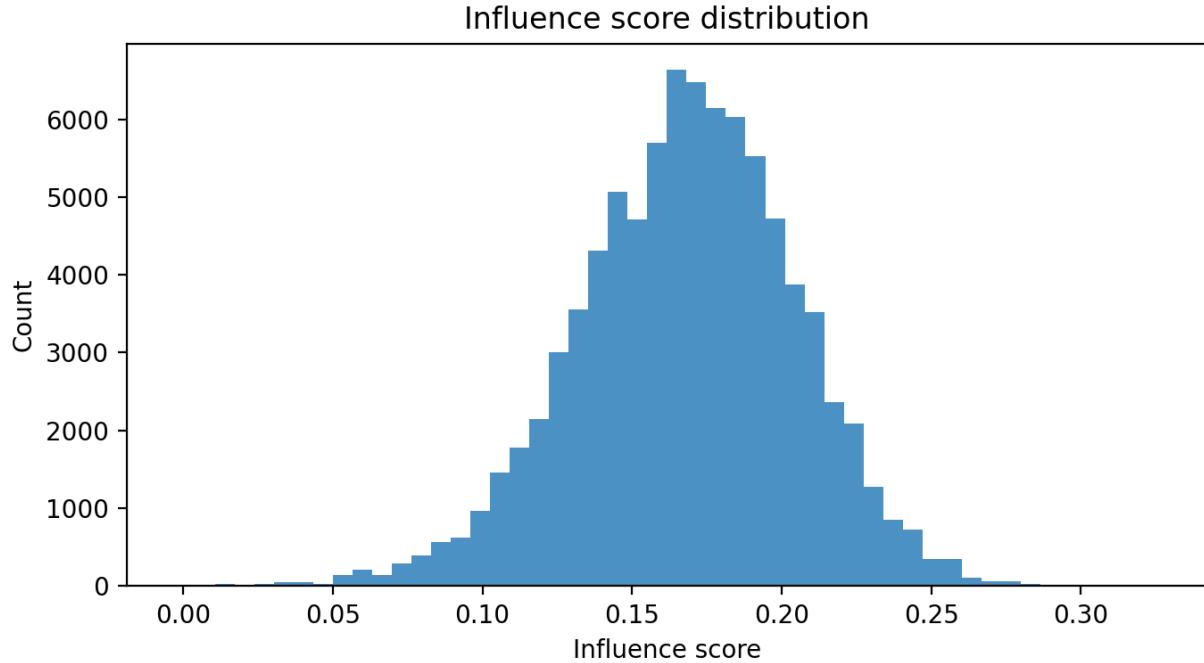
Figure 2. Distribution of per-sample TracIn influence scores. Influence values cluster near zero with a long positive tail and a small negative region, reflecting focal loss–driven gradient amplification on difficult or mislabeled examples.

2.5 Subset Experiments

Two experiments assess dataset redundancy and compressibility. In the removal experiment, the top-influence samples are removed before retraining the model to determine how much performance depends on highly influential data points. In the selection experiment, the model is trained only on the samples with the highest influence scores to determine how much of the dataset is required to recover baseline performance. As a whole, these experiments reveal how learning is distributed across the dataset.

3   Results

3.1 Baseline Performance

The DenseNet-121 model trained under focal loss achieves a macro AUROC of 0.8139, a macro F1 score of 0.3502, and an expected calibration error (ECE) of 0.2348. Per-class AUROC varies substantially. Well-defined conditions such as emphysema and hernia achieve scores above 0.90, whereas ambiguous conditions such as infiltration and pneumonia score

notably lower (~0.70). These disparities reflect the label noise and conceptual ambiguity previously documented in NIH14 (Oakden-Rayner, 2017).

3.2 Influence Distribution

Influence scores range from −0.00246 to 0.32580. The distribution is long-tailed with most values clustered near zero, consistent with TracIn's behavior under focal loss. The presence of negative influence values indicates that certain samples consistently worsen validation loss, a phenomenon often associated with mislabeled or clinically atypical images.

3.3 Removal Experiments

Removing up to 35% of the highest-influence samples produces only minimal performance degradation. Across removal levels from 5% to 35%, AUROC remains between 0.8068 and 0.8139, and macro F1 remains stable as well. Performance declines more noticeably only when 60% of the most influential samples are removed. These findings suggest that NIH14 contains substantial redundancy and that many high-influence samples may not be beneficial for generalization.
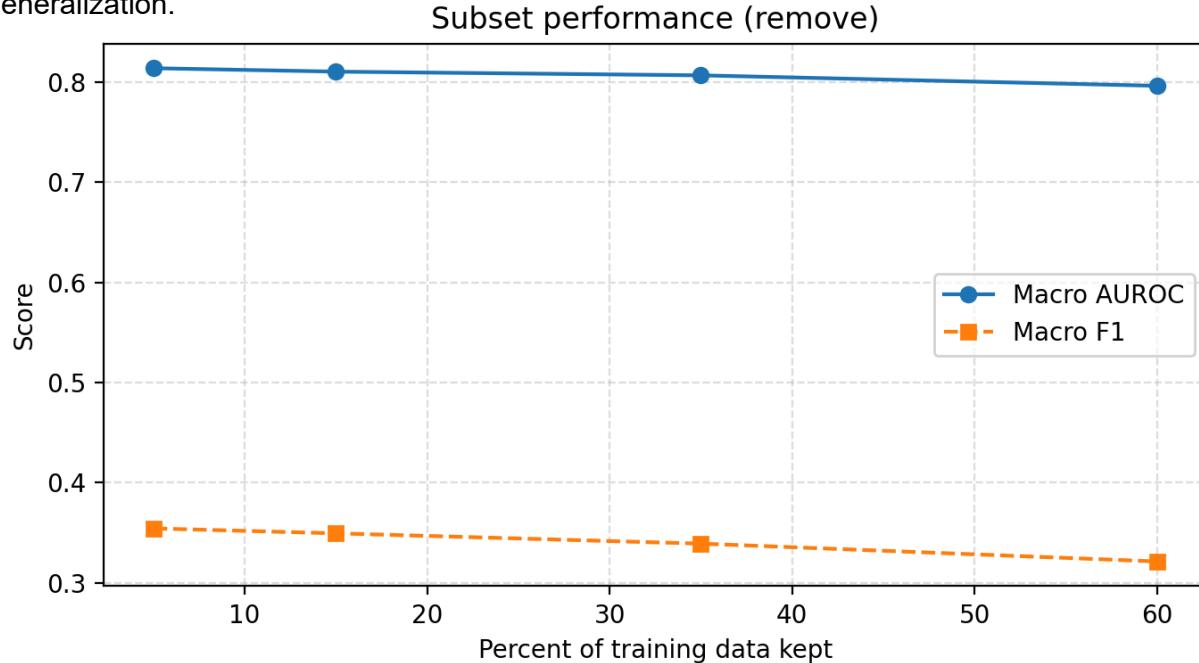
Figure 3. Performance curves showing the effect of removing top-influence samples. AUROC and F1 remain stable for removals up to 35%, demonstrating strong dataset redundancy.

3.4 Selection Experiments

When training solely on top-influence samples, performance is initially poor at extremely small subset sizes but improves rapidly as more samples are included. With only 10% of the dataset, the model achieves a macro AUROC of 0.7194. Using 40% yields 0.7858, and using 80% nearly matches baseline performance. These results indicate that a relatively small fraction of samples carries most of the learning signal, supporting the idea that NIH14 is highly compressible
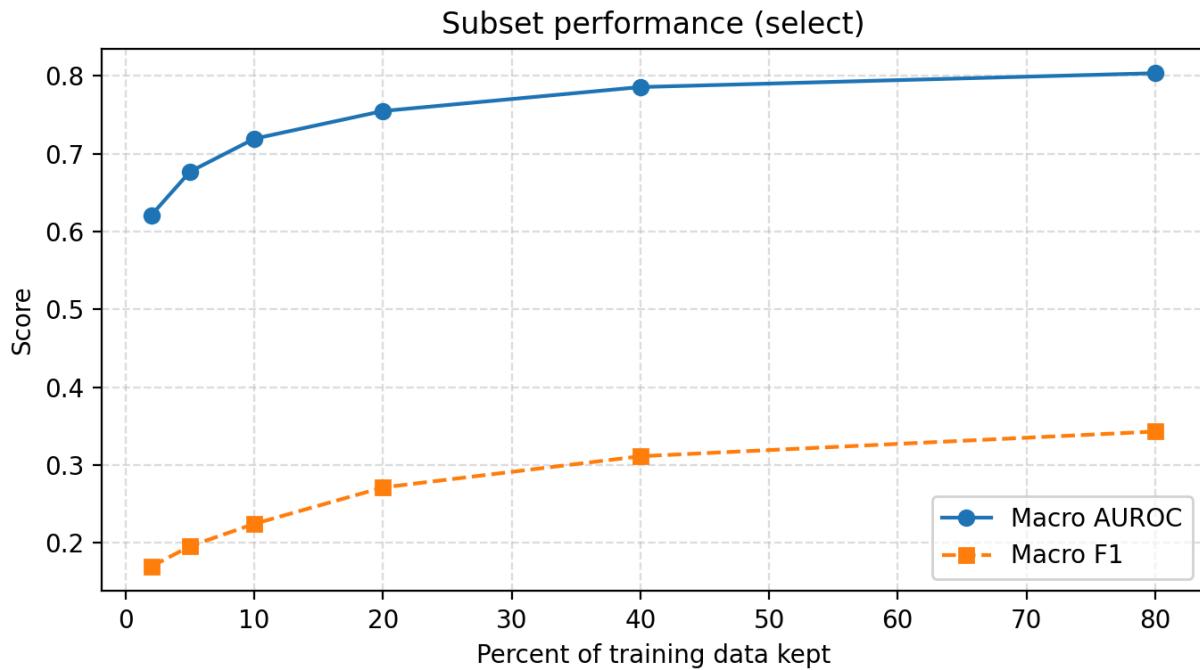
Figure 4. Performance of models trained only on the highest-influence samples. Performance recovers rapidly as subset size increases, indicating strong dataset compressibility.

4   Discussion

The influence analyses reveal substantial redundancy and noise within NIH14. The removal experiments demonstrate that even samples with the highest positive influence can often be removed without harming performance, a pattern consistent with earlier critiques of NIH14's noisy and weakly defined labels (Oakden-Rayner, 2017). This suggests that influence magnitude is not necessarily synonymous with usefulness, and that the dataset contains many samples that are influential but not informative. The selection experiments further show that training on a relatively small subset of samples recovers most of the full-dataset performance, implying that the dataset's effective information content is far smaller than its nominal size. This

aligns with the broader data-centric AI argument that model performance is constrained primarily by dataset quality rather than architectural capacity (Shankar et al., 2023).

These observations also resonate with concerns raised in human-centered ML, which emphasizes the sociotechnical nature of dataset construction and the ways in which noisy or unstable labels can propagate into downstream system behavior (Wang et al., 2017; Oakden-Rayner, 2017). In this context, highly influential yet unreliable samples may steer optimization toward patterns that reflect dataset artifacts rather than clinical truth. Influence scoring provides a practical means of diagnosing such issues: the presence of negative-influence samples identifies training cases that actively degrade validation performance, providing clear targets for clinician review, dataset refinement, or improved annotation guidelines.

Influence-based methods like TracIn (Pruthi et al., 2020), building on the theoretical foundations of influence functions (Koh & Liang, 2017), therefore serve not only as interpretability tools but also as mechanisms for responsible dataset governance. Their ability to surface mislabeled, atypical, or overly dominant samples directly supports human-in-the-loop curation workflows advocated in HCML frameworks (Shneiderman, 2020). Future work could extend these findings by analyzing influence stability across training epochs, integrating uncertainty-aware valuation methods, or conducting clinician-guided review of extreme-influence samples to better understand their pathological characteristics and inform targeted dataset improvement strategies.

5 Conclusion

This project demonstrates that influence-based sample valuation provides a powerful lens for understanding dataset structure in noisy clinical imaging tasks. Using TracIn influence scoring and focal-loss training, the analysis reveals that NIH ChestXray14 contains extensive redundancy and a relatively small core of samples that drive most of the model's learning. These findings resonate with data-centric AI principles and emphasize the importance of dataset curation, label quality, and interpretability in high-stakes medical applications. Influence-guided auditing offers a promising path toward more reliable, human-centered AI systems in healthcare.

References

Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. ICML.
 Pruthi, D., Liu, F., Kale, S., & Sundararajan, M. (2020). Estimating Training Data Influence by Tracing Gradient Descent. NeurIPS.

Wang, X. et al. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks.
Oakden-Rayner, L. (2017). Exploring the ChestXray14 dataset.
Shankar, S. et al. (2023). DataPerf: Benchmarking Data Quality for Performance in Machine Learning.
Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human–Computer Interaction.

Link to Github Repo: https://github.com/danialressl/CMPT_419_Final_Project-