**Enhancing Chest X-ray Pathology Classification with a Rejection Mechanism**

**Amit Tzahar, Alon Gottlib, Tal Verber, Ravit Shagan Damti**

**Abstract:** This paper presents an enhancement to a chest X-ray pathology classification model by integrating a rejection mechanism. The base model, implemented as described in the paper [*Multi-Domain Balanced Sampling Improves Out-of-Distribution Generalization of Chest X-ray Pathology Prediction Models*](https://www.cse.cuhk.edu.hk/~qdou/public/medneurips2021/77_chest_ood_med_neurips_2021.pdf), classifies images into four pathology categories. We introduce a rejection mechanism that identifies uncertain predictions and discards them, thus aiming to improve the reliability of the model. Two rejection methods were compared, with experimental results that demonstrate the trade-off between rejection rate and classification accuracy.

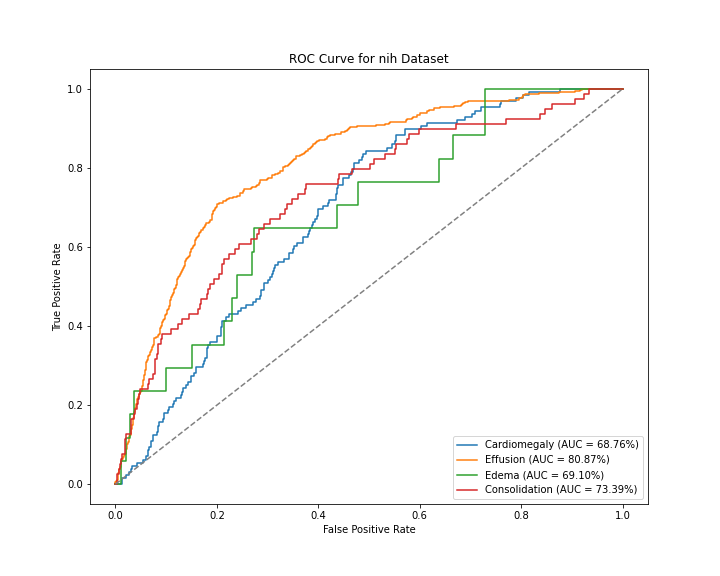
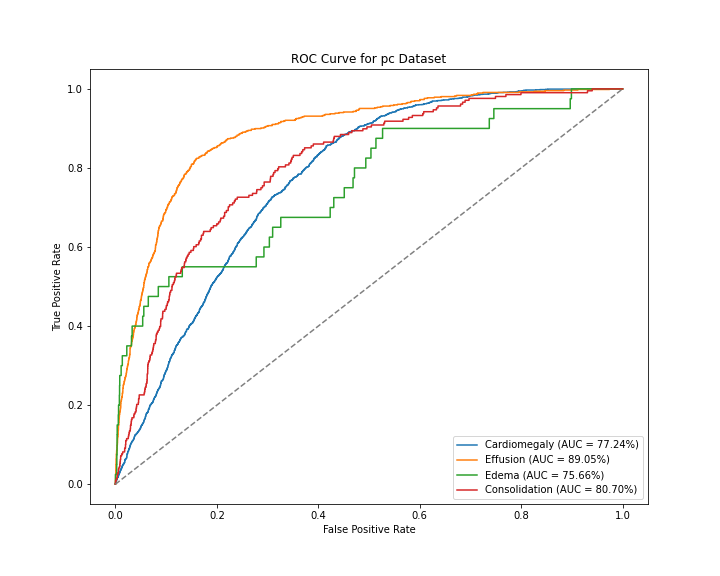
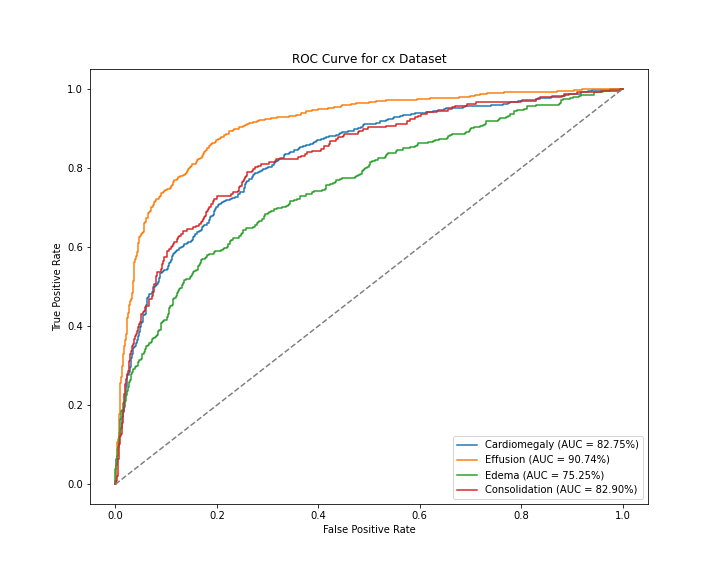
You can find the full code on our [github](https://github.com/TzaharAmit/X-Ray-Pathology-Prediction-w-Rejection).

**1. Problem Description**: Automated medical image classification models are susceptible to errors, especially when faced with out-of-distribution (OOD) data. A model making incorrect predictions on uncertain samples may lead to misdiagnoses, which can have serious consequences in clinical settings. Implementing a rejection mechanism allows the model to abstain from making uncertain predictions, thereby improving overall reliability and ensuring that only confident predictions are retained.

**2. Dataset Description**: The dataset used consists of chest X-ray images with labeled pathologies. The classification model categorizes each image into one or more of four conditions: Cardiomegaly, Effusion, Edema, and Consolidation. Unlike traditional single-label classification tasks, a single image may belong to multiple pathology categories simultaneously. The dataset includes multiple sources to improve generalization and was split into training, validation, and test sets. For each image, a confidence score was extracted from the model, which served as the basis for the rejection mechanism. Table 1 presents the distribution of each class:

|  |  |  |
| --- | --- | --- |
| **Category** | **Positive Count** | **Percentage (%)** |
| Cardiomegaly | 11919 | 9.70366 |
| Effusion | 9353 | 7.61459 |
| Edema | 1956 | 1.59244 |
| Consolidation | 2650 | 2.15745 |
| Multi-Pathology | 3869 | 3.14988 |
| Total Samples | 122830 | 100 |

**3. Baseline Model Description**: The baseline model follows a deep learning-based approach (DenseNet-121) for pathology classification. It was pre-trained on large-scale chest X-ray datasets and fine-tuned using a multi-domain balanced sampling strategy to improve generalization to unseen data. The model outputs a probability score for each pathology. The AUC of the model was calculated to understand the baseline results.



*Figure 1: Basline Model Results*

**4. Rejection Mechanisms**: The rejection mechanism works after the model run is finished. It works on the output files (a file for each data set) of the model that ran, which contain probabilities for each pathology (as mentioned earlier, each sample can belong to several pathologies, so each pathology has a probability between 0 and 1).

Several rejection strategies were considered and compared:

**4.1 Entropy-based Rejection**

We evaluated two different data‐splitting strategies:

1. **Within‐Dataset Split:** Each dataset is split 80% for training and 20% for validation, keeping train/validation data from the same distribution.
2. **Cross‐Dataset Split:** Two entire datasets are used for training, and the third dataset is reserved for validation. This simulates an out‐of‐distribution scenario, since the validation set may differ significantly from the training distribution.

The following analysis focuses on the first approach (the 80%–20% within‐dataset split).

Since each image can have more than one pathology, this means the task is a **multi-label classification** problem. This fact directly affects the results of the rejection mechanism.

1. Rejection Can Discard Too Many Samples

Each image in the dataset can have multiple pathologies. By applying rejection on a per-pathology basis, an image might be rejected because of one pathology, even though it was correctly classified for another pathology.

This causes a loss of useful data, leading to a larger drop in AUC after rejection.

1. Rejection Rate is Amplified in Multi-Label Cases

In multi-label classification, If the rejection is applied independently for each pathology, it can lead to over-rejection.

Our approach to handle the multi-label classification:

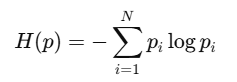
1. Apply Rejection Based on the Entire Image, Not Per-Pathology

Instead of rejecting each pathology separately, reject an image only if all predictions for that image are below entropy thresholds. This ensures that images with at least one confident prediction are kept.

1. Use a More Adaptive Rejection Threshold

Instead of fixed Entropy thresholds, we Compute a per-pathology adaptive threshold based on the training set distribution. We set the threshold dynamically using the percentile of correctly classified samples' scores.

**Shannon Entropy** measures uncertainty in a probability distribution. For a single prediction with N classes, it’s defined as:



In a **binary** setting (each pathology is either present or absent), the entropy for a single pathology is:



where p is the predicted probability for that pathology.

**Maximum possible entropy** for a single probability p occurs at p=0.5, which gives:

  
That’s the “peak uncertainty”, since the model is equally likely to predict p=0.5 for present vs. absent.

Entropy Threshold meaning intuition:

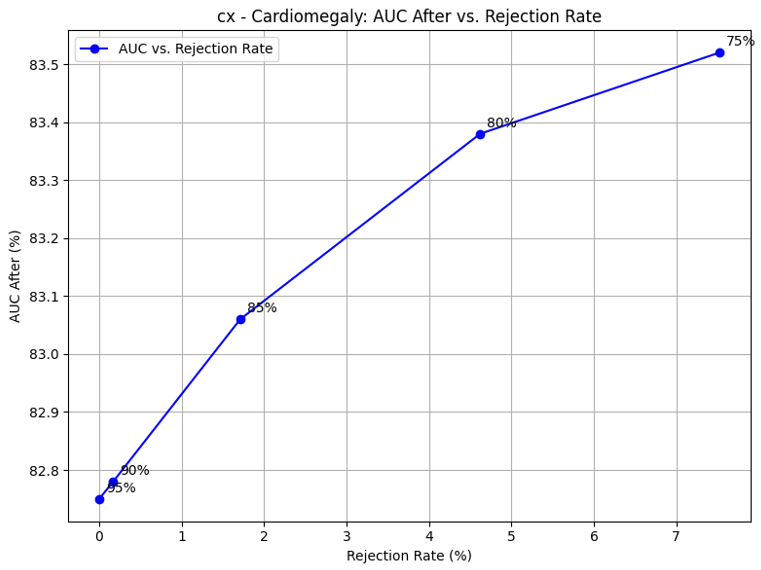
A threshold of 0.65 is therefore close to the maximum (0.6931).

* Any predicted probability near 0.5 will have an entropy near or above 0.65.
* So if the entropy for a sample exceeds **0.65**, it indicates the model is “too uncertain” about that pathology’s presence or absence.

Entropy Threshold Optimization process:

1. Select a range of candidate percentiles between 75% and 95%.
2. For each pathology, compute an adaptive entropy threshold using the chosen percentile from the distribution of entropy values of correctly classified samples in the training set (thresholds are computed separately for each pathology).
3. Evaluate the impact of each percentile by plotting the resulting AUC and rejection rate.
4. Choose the percentile that achieves a strong AUC while minimizing the rejection rate (here the Rejection rate was limited to 25% max. the maximum value can be adjusted per doctor request)

Example (CX dataset, Cardimegaly):



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Pathology** | **Percentile** | **Entropy TH** | **AUC Before** | **AUC After** | **Rejection Rate (%)** |
| cx | Cardiomegaly | 75 | 0.6559 | 84.54 | 84.98 | 6.71 |

1. Calculate the AVG Optimal Percentile across 3 different Datasets for each pathology.
2. Set this Percentile = Set the TH (Pathology Entropy TH)

Test the model after adding the rejection mechanism:

On the Validation set - run the Rejection mechanism with the optimal TH values.

Entropy Threshold Optimization process - Outcome:

The optimal percentile per pathology we used for the test:

|  |  |
| --- | --- |
| **Pathology** | **Average Optimal Percentile** |
| Cardiomegaly | 76.3 |
| Consolidation | 83.6 |
| Edema | 81 |
| Effusion | 86.3 |

Results

Validation set AUC results using Optimized Entropy TH:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Pathology** | **Entropy TH** | **AUC Before** | **AUC After** | **F1 Before (%)** | **F1 After (%)** | **Rejection Rate (%)** |
| pc | Cardiomegaly | 0.5886 | 77.24 | 78.1 | 30.38 | 28.98 | 15.53 |
| Effusion | 0.6239 | 89.05 | 88.61 | 29.78 | 26.61 | 35.98 |
| Edema | 0.6904 | 75.66 | 78.78 | 0.64 | 0.6 | 15 |
| Consolidation | 0.63 | 80.7 | 80.88 | 6.71 | 6.43 | 25 |
| nih | Cardiomegaly | 0.583 | 68.76 | 70.93 | 9.99 | 8.89 | 25 |
| Effusion | 0.6388 | 80.87 | 78.61 | 36.7 | 29.05 | 45.73 |
| Edema | 0.6847 | 69.1 | 67.94 | 1.19 | 0.98 | 29.41 |
| Consolidation | 0.6489 | 73.39 | 68.04 | 7.8 | 5.05 | 43.04 |
| cx | Cardiomegaly | 0.6592 | 82.75 | 83.22 | 61.95 | 62.78 | 2.91 |
| Effusion | 0.6911 | 90.74 | 91.11 | 79.67 | 80.08 | 3.03 |
| Edema | 0.6924 | 75.25 | 75.36 | 33.84 | 34.6 | 0.88 |
| Consolidation | 0.6922 | 82.9 | 83.01 | 22.8 | 23.07 | 1.69 |

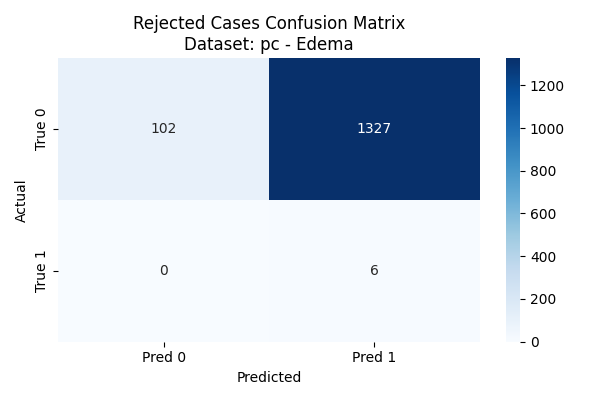
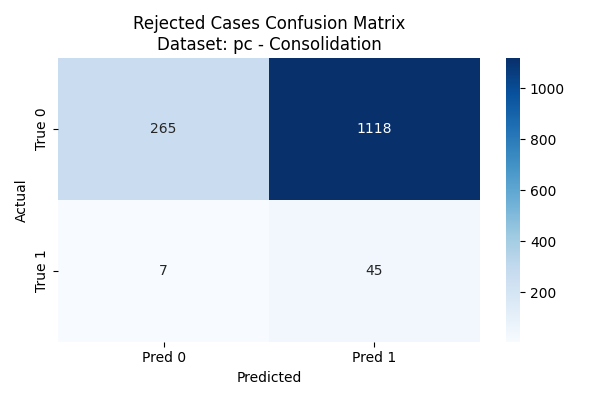
Findings:

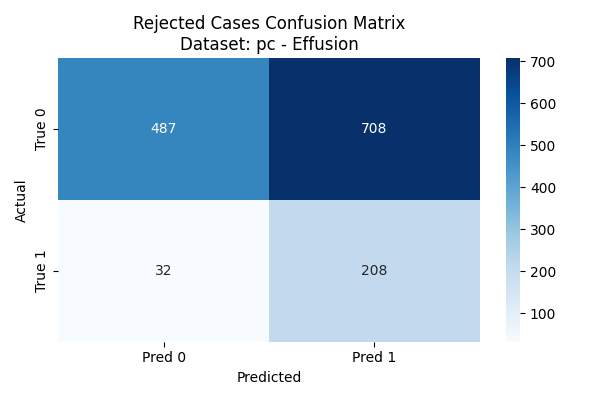
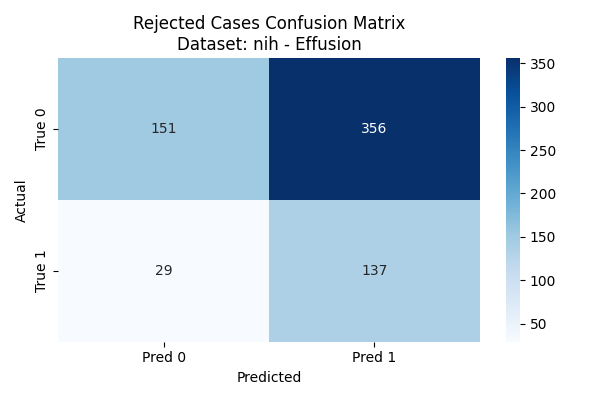
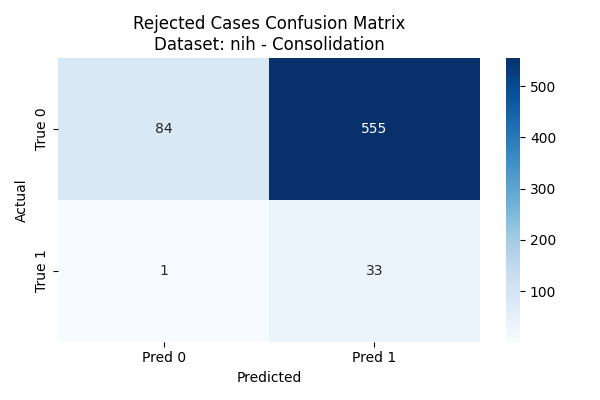
* Entropy-based rejection didn’t consistently improve AUC or F1. In some cases, these metrics dropped.
* Possible Reason: The percentile threshold (~0.65–0.69) can remove many borderline but correct samples, hurting performance.
* High Rejection Rate in certain pathologies suggests the threshold is too strict, especially if the model is miscalibrated. In a miscalibrated model, an entropy value might not truly reflect the model's uncertainty.
* Limiting the Rejection Rate to max value for the training set (during TH optimization process) does not guarantee that the rejection rate for the validation set will be below this value. once we pick a percentile from the filtered training results and apply it on the validation set, the resulting thresholds can still yield a rejection rate above 25% on the validation data, because:
  + Different Data Distributions -The validation set can have a different entropy distribution compared to the merged training set.
  + We Are Filtering on Training Rejection Only - it’s entirely possible for the chosen threshold to produce a higher rejection rate on validation.

In order to guarantee that both training and validation rejection rates do not exceed 25%, we need to change the optimization process to test each percentile on both sets before accepting it.

## Confusion Matrices for Rejected Validation Samples

Below are selected examples of the resulting confusion matrices:



## These matrices show only the images that were rejected (the model deemed them uncertain enough under the “all-above-threshold” rule). The cells (TN, FP, FN, TP) reflect how those rejected samples break down relative to ground truth vs. predicted labels (using a 0.5 cutoff).

Overall, the Rejection Mechanism Successfully Removes Many Incorrect Positives - The large FP counts suggest the rejection mechanism is doing its job in filtering out borderline, misclassified images.

## Common Patterns Across All 12 Matrices:

1. **Large Blocks of False Positives (FP)**

In many pathologies (e.g., pc\_Edema with 1327 FPs, pc\_Consolidation with 1118 FPs, nih\_Consolidation with 555 FPs, etc.), we see very high counts of false positives among the rejected set.

This indicates that many borderline (or uncertain) images get predicted “1” but are actually “0.” The rejection mechanism removes a lot of these incorrect positives, which is typically beneficial for overall precision.

1. **Some True Positives (TP) Among Rejections**

Some truly positive images appear in the rejected group. For instance, pc\_Effusion shows 208 TPs among the rejected set, nih\_Effusion has 137 TPs, etc.

This implies that not all rejections are purely mistakes; some correct positives are also being discarded because the entire image is considered “uncertain” for the other pathologies.

1. **Few True Negatives (TN) or False Negatives (FN) in Some Cases**

In certain pathologies (e.g., cx\_Consolidation or cx\_Edema), we see zero or very few TN or FN in the rejected set. Often the model predicted “1” for negative samples, leading to large FP blocks, while truly negative images predicted negative seldom end up in the rejected set.

Some pathologies show a moderate number of FNs (e.g., pc\_Effusion with 32 FNs), indicating that some positives were missed altogether in those uncertain images.

1. **Dataset Differences**

The **pc** dataset also discards hundreds of images for certain pathologies (e.g., 1118 FPs for Consolidation, 1327 for Edema), similar to **nih**.

The **cx** dataset often has fewer rejections or smaller numbers, but the ratio of FP to TP can still be quite large.

Overall, each dataset sees a significant chunk of borderline images flagged as uncertain, removing many false positives but also losing some correct positives.

**4.2 Based on a confidence interval** (In this strategy each output file is divided into 80% for training and 20% for validation):

The selection of the optimal rejection threshold (TH) was performed in two stages:

1. Finding the Optimal Threshold per Pathology in Each Dataset:
   * For each pathology, we computed the probability distribution of correctly classified samples.
   * The standard deviation of the probability distribution was used as a key metric to define variability.
   * The optimal TH was determined by analyzing the median confidence score and computing absolute deviations from the median.
   * Depending on the skewness of the distribution, different statistical techniques were applied (e.g., trimmed mean and weighted variance adjustments).
   * A weighted scoring method was used to balance different evaluation metrics when selecting the threshold according to **Skewness.**
2. Selecting the Final TH for Each Pathology:
   * The optimal thresholds found per dataset were aggregated.
   * A final threshold was selected by computing the average across all datasets, ensuring stability and robustness.
   * This ensured consistency across different data sources and improved generalization.

Table 2: Rejection Rate:

| **Category Name** | **Threshold** | **Rejection Rate** | **F1 Score** | **Weighted Score** |
| --- | --- | --- | --- | --- |
| Cardiomegaly | 0.0329 | 0.09 | 0.227 | 0.345 |
| Effusion | 0.0228 | 0.07 | 0.275 | 0.498 |
| Edema | 0.0288 | 0.23 | 0.126 | 0.178 |
| Consolidation | 0.0268 | 0.27 | 0.143 | 0.199 |

Rejection Decision Logic:

* The rejection mechanism was based on the probability distribution, standard deviation, and median deviation of predictions.
* **Right-skewed distribution (Skewness > 1.0):** Reject samples with low confidence scores, i.e., probability values below the median minus a multiple of the standard deviation.
* **Left-skewed distribution (Skewness < -1.0):** Reject samples with high confidence scores that deviate significantly above the median.
* **Symmetric distribution (-1.0 ≤ Skewness ≤ 1.0):** Reject samples where the absolute deviation from the median exceeds the predefined threshold.
* The rejection decision dynamically adapts to the dataset’s statistical properties, ensuring optimal performance across different conditions.

**5. Model Parameters and Implementation**: The model was implemented using PyTorch and trained on a multi-GPU system. The rejection mechanism was integrated into the inference pipeline. The key parameters included:

**5.1 Confidence-based Rejection**

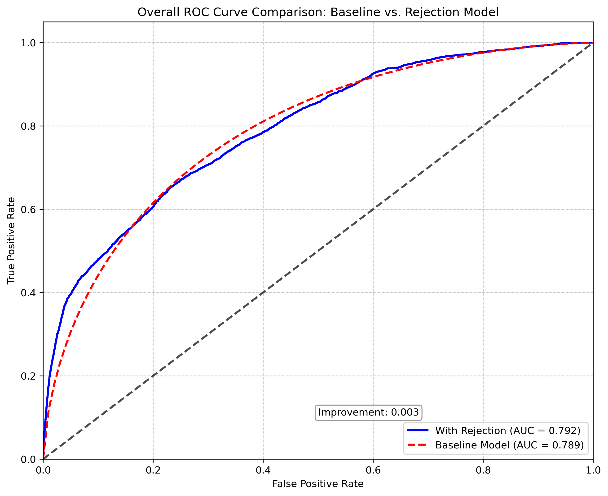
* **Probability Threshold (T):** Determines when a sample is rejected.
* **Standard Deviation-Based Variability Measure:** Used to determine rejection criteria dynamically.
* **Skewness Threshold:** Identifies highly skewed distributions requiring different rejection criteria.
* **Statistical Features:** Median, trimmed variance, and histogram-based measures were computed to refine rejection decisions.
* **Evaluation Metrics:** Precision, recall, F1-score, AUC, and rejection rate were measured.

**5.2 Entropy-based Rejection**

* **Entropy Threshold (T):** Shannon entropy calculation of the predicted probability. Setting a threshold based on a chosen percentile of the entropy values from correctly classified samples. This threshold (T) is used to decide whether a given prediction is too uncertain.
* **Percentile parameter:** A tunable parameter that determines the entropy threshold. A lower percentile will produce a lower threshold (making the mechanism more aggressive, i.e., rejecting more samples), while a higher percentile results in a higher threshold (more lenient rejection).
* **Evaluation Metrics:** Standard metrics such as AUC, F1 Score, and Rejection Rate were measured.

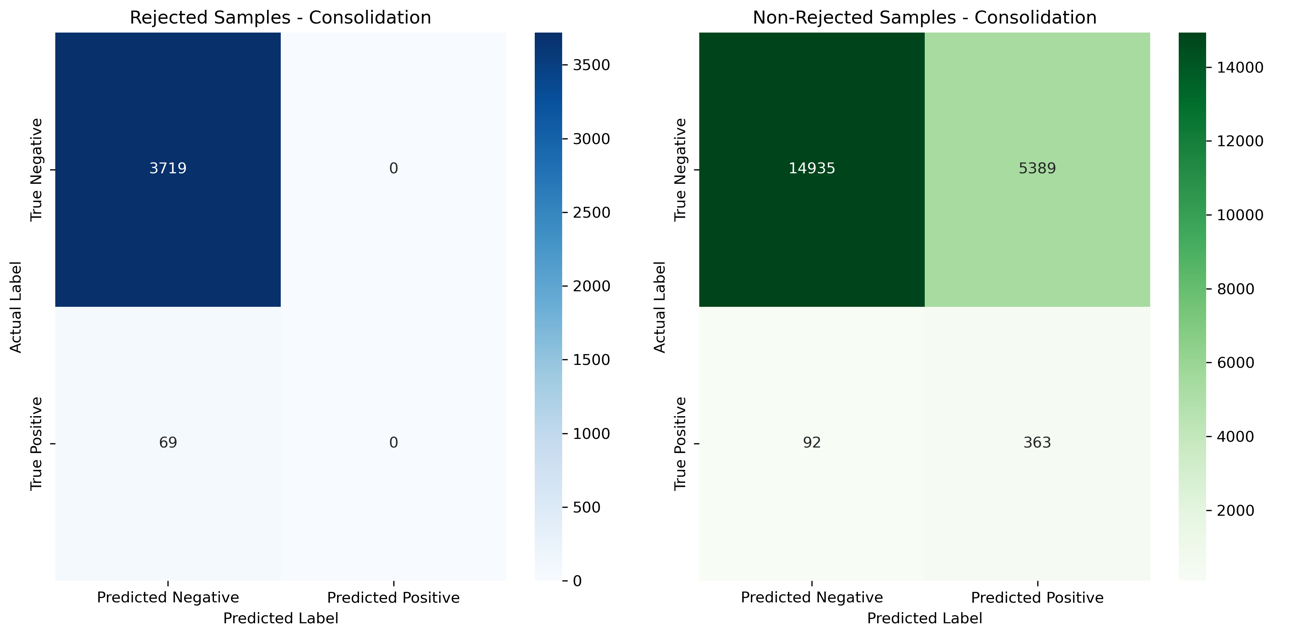
**6. Experiments**: The model was evaluated with and without the rejection mechanism on multiple datasets. The experiments tested different rejection strategies and their impact on classification performance.

* The optimal rejection threshold was determined through statistical analysis of model outputs.
* Performance was assessed in terms of classification accuracy, rejection rate, and retained dataset quality.
* AUC and precision-recall curves were used to visualize the trade-offs.



**7. Results**: The introduction of the rejection mechanism improved the reliability of the model:

* Accuracy vs. Rejection Rate: As rejection increased, accuracy improved on the retained samples.
* Comparison of Rejection Strategies: The margin-based approach outperformed simple thresholding, while adaptive rejection yielded the best balance between accuracy and rejection rate.
* Effect on AUC: The model with rejection outperformed the baseline model in terms of AUC, particularly on difficult cases.
* Confusion Matrix Analysis: The rejection mechanism successfully filtered out samples that would have led to misclassifications.

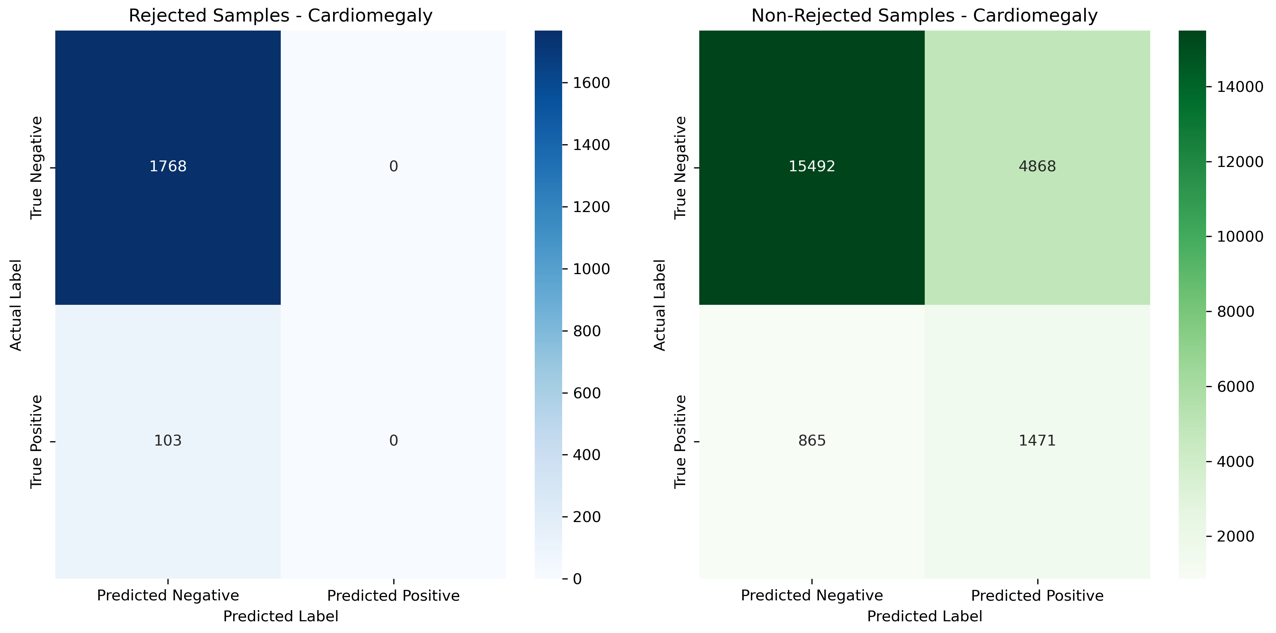


תמונה שמכילה טקסט, צילום מסך, תרשים, מספר

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

תמונה שמכילה טקסט, צילום מסך, תרשים, מספר

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.



**Results – Both Methods Comparison:**

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset  Category | Mean AUC – Baseline Model | Mean AUC – Confidence Rejection | Mean AUC – Entropy Rejection |
| Cardiomegaly | 0.76 | 0.76 | 0.77 |
| Effusion | 0.79 | 0.87 | 0.86 |
| Edema | 0.73 | 0.73 | 0.74 |
| Consolidation | 0.87 | 0.79 | 0.77 |
| Avg Test AUC | 0.79 | 0.79 | 0.79 |

1. **Conclusion and Future Work:**

* Rejection mechanisms can enhance reliability by discarding low-confidence or high-entropy predictions—but only when thresholds are carefully tuned.
* In multi-label classification, rejecting entire samples based on all pathologies may discard valuable correct predictions, a risk you mitigated with an "all-below-threshold" rejection rule.
* Adaptive thresholds (based on percentiles) are more robust than fixed ones, but differences between training and validation distributions can lead to unexpectedly high rejection rates.
* The confusion matrix analysis showed that our rejection mechanism effectively removed many false positives, improving model precision.
* There’s no “one-size-fits-all” strategy - different pathologies behave differently under rejection mechanisms.

Future research could focus on analyzing the impact of rejection mechanisms on specific cases to determine whether rejected samples are truly uncertain or misclassified. Investigating the characteristics of these samples can provide insights into refining rejection strategies for better decision-making. Additionally, exploring hybrid rejection approaches that combine confidence-based and entropy-based methods could enhance overall performance by leveraging the strengths of both techniques—filtering low-confidence predictions while capturing distributional uncertainty. Furthermore, extending rejection mechanisms to multi-label classification for further study, as handling uncertainty in multi-target scenarios presents unique challenges. Developing tailored strategies for multi-label settings could improve the reliability and interpretability of rejection-based decision systems. Also, tuning the rejection threshold by experimenting with different percentiles to balance the removal of truly incorrect samples with the retention of correct borderline ones, and ensuring the rejection rate remains acceptable by constraining it during the threshold optimization process. Additionally, improving model calibration is crucial so that entropy more accurately reflects prediction correctness, as a well-calibrated model ensures that predicted probabilities—and thus entropy—align with actual correctness likelihoods.