Feature Importance

Image Importance Score: 28.72800064086914

Metadata Feature Importance Scores: {'age_approx': 0.003150713862851262, 'clin_size_long_diam_mm': 0.0004841720510739833, 'sex_female': 7.600174285471439e-07, 'sex_male': 0.00032329559326171875, 'anatom_site_general_anterior torso': 6.821283022873104e-07, 'anatom_site_general_head/neck': 0.0, 'anatom_site_general_lower extremity': -0.0012100040912628174, 'anatom_site_general_posterior torso': 0.0, 'anatom_site_general_upper extremity': 0.0}

Image Importance

The Image Importance Score is significantly higher compared to the metadata importance scores, with a value of 28.728. This indicates that the model heavily relies on image data to make predictions. Given that skin lesion classification is a visual task, this result aligns with expectations, as the image contains the most direct and relevant information about lesion characteristics.

Metadata Importance

The Metadata Feature Importance Scores are comparatively very low, suggesting that the metadata contributes minimally to the model's decision-making. However, some features stand out more than others:

- **age_approx (0.0031)**: This is the most influential metadata feature. This could be because certain types of lesions are more prevalent in specific age groups, providing the model with some additional context.

- **clin_size_long_diam_mm (0.00048)**: The clinical size of the lesion contributes to predictions, albeit minimally. Lesion size may be relevant for distinguishing between benign and malignant cases but is overshadowed by the image data.

- **sex_male (0.00032)** and **sex_female (close to zero)**: The gender-related features contribute very little, indicating that the model does not rely heavily on sex to differentiate between lesion types.

- Some features like **anatom_site_general_lower extremity (-0.0012)** exhibit negative importance. Negative values might indicate that removing this feature slightly improves model predictions, possibly due to noise or redundancy in the data. Other **anatomical sites like head/neck, posterior torso, and upper extremity** show zero importance, suggesting they are not used by the model.

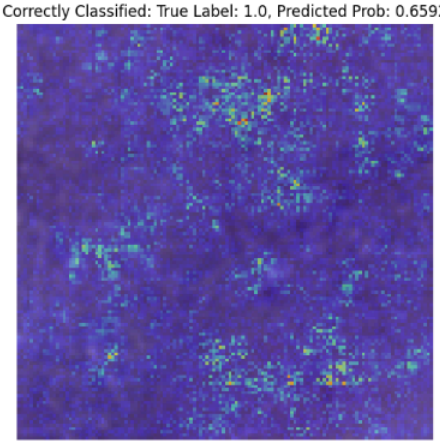
To Improve Metadata Contribution, I would like to try to explore metadata-only models or ensemble approaches to amplify the metadata's role in predictions.

5 Predictions

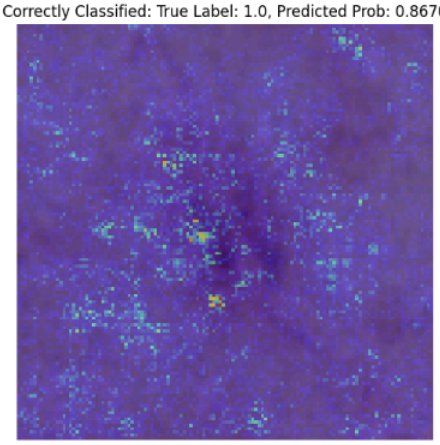
Correctly Classified

Displaying correctly classified samples with true label 1:

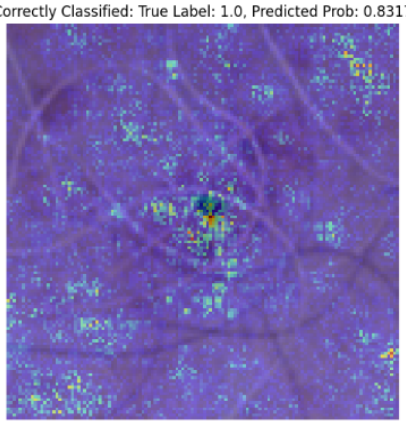
Correctly Classified: True Label: 1.0, Predicted Prob: 0.6592



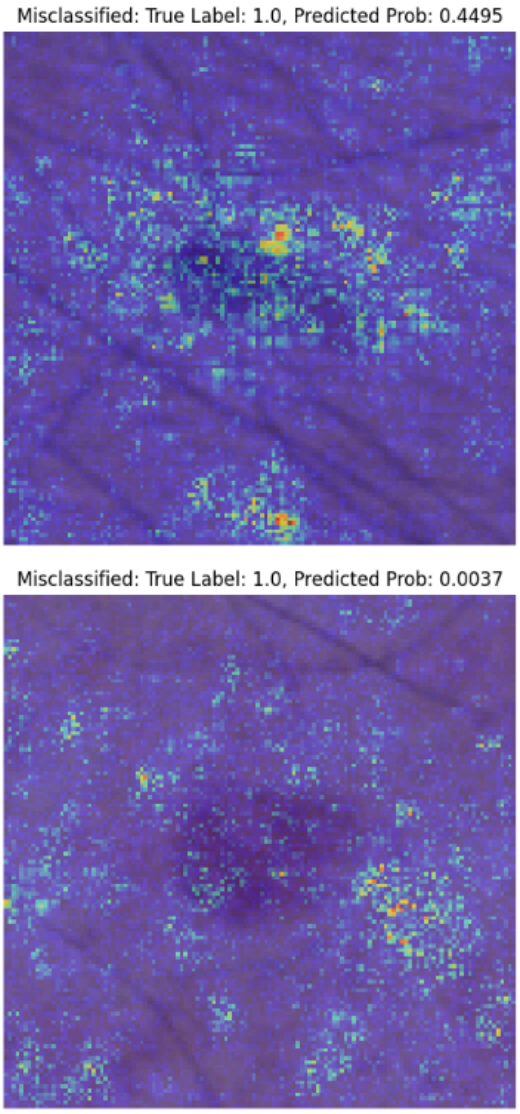
Correctly Classified: True Label: 1.0, Predicted Prob: 0.8670



Correctly Classified: True Label: 1.0, Predicted Prob: 0.8317



Misclassified



After analyzing the feature importance, I shifted my focus to understanding how the model classifies class 1 images, which represent malignant skin lesions. These cases are critical as they directly impact the effectiveness of the model in identifying potentially life-threatening conditions.

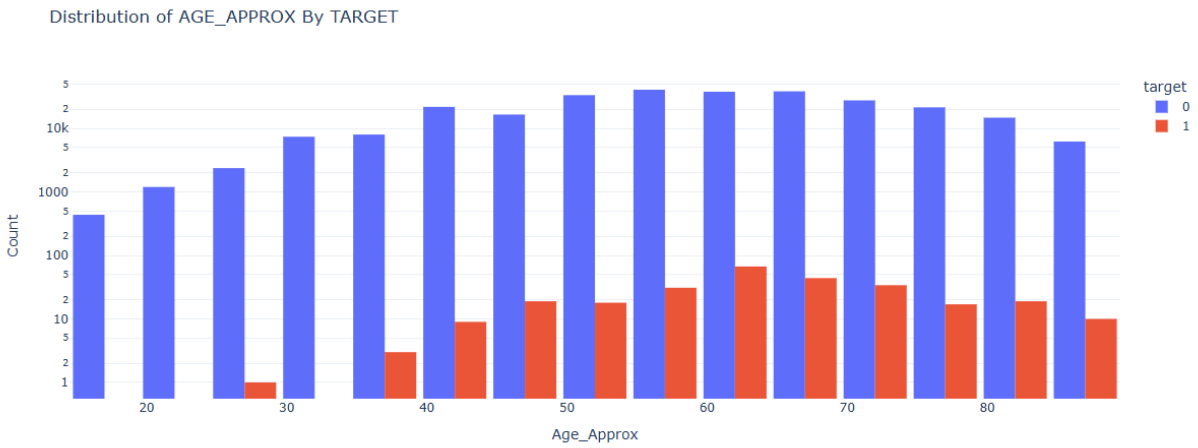
To gain deeper insights, I employed heatmaps to visualize the regions of the images that contributed the most to the model's predictions. Heatmaps allow us to assess whether the model is focusing on the relevant areas of the lesion, such as texture, color variations, or edges, which are clinically significant features for diagnosis. By overlaying the heatmap on the original image, we can determine if the model is attending to the lesion itself or if it is

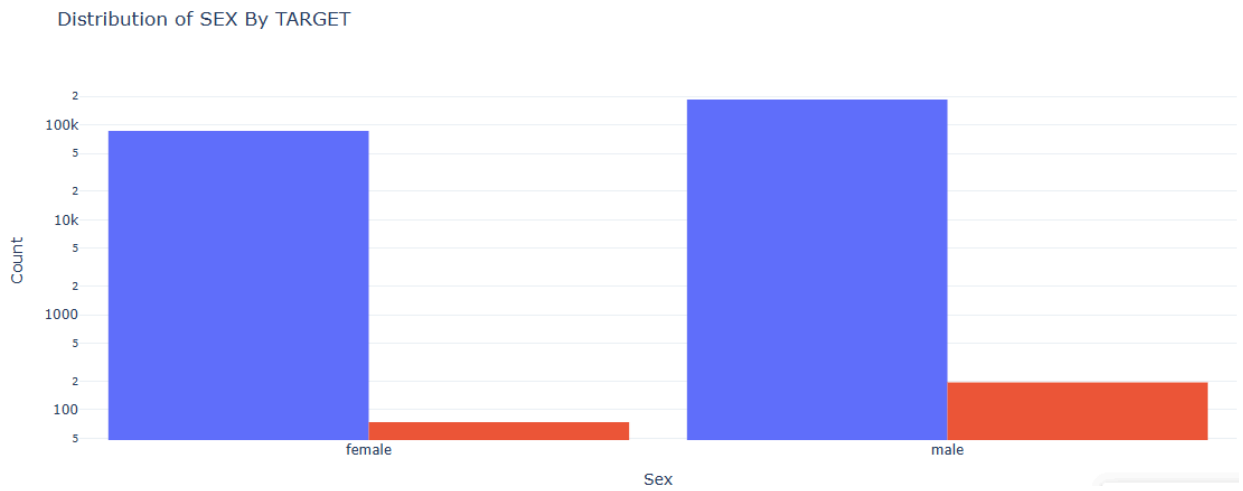
being influenced by irrelevant factors, such as the background or artifacts. From the images above, it is evident that the misclassified images are primarily due to the model focusing on areas far from the actual skin lesion. This indicates that the model's attention is not centered on the region of interest. To address this issue, I propose incorporating a masking technique as an additional input. Masking will help the model focus on the skin lesion area by explicitly highlighting the region of interest. This could improve the model's ability to differentiate between benign and malignant lesions, as it would reduce distractions from irrelevant parts of the image and ensure the model learns features directly related to the lesion. Incorporating a mask could involve preprocessing steps to identify and isolate the lesion, followed by feeding this processed data alongside the original image and metadata into the model.

This analysis provides valuable information about the model's interpretability and helps identify potential biases or weaknesses. For example, if the heatmap highlights areas outside the lesion, it may indicate that the model relies on non-relevant features, which could lead to misclassification. By understanding these patterns, we can refine the model further to improve its focus and accuracy for classifying malignant skin lesions.

Protected Categories

The dataset contains age and sex as protected categories, which are used in the model as part of the metadata input.





Bias of the Model

Yes, the dataset and potentially the model output exhibit bias due to:

- **Class Imbalance:** There are significantly fewer malignant cases (target = 1) compared to non-malignant cases (target = 0) in both the age and sex distributions.
- **Protected Classes:** Age groups under 40 and females, in particular, are underrepresented for malignant cases. This could lead to lower recall for these subgroups, as the model may not learn enough from the available data.

Bias removal strategies

Data preprocessing strategies

Balancing data through techniques like oversampling underrepresented groups or undersampling overrepresented groups can help create a more equitable dataset. Another approach is feature censoring, where sensitive attributes (e.g., gender or age) are removed entirely to prevent the model from learning their effects. Fair data generation methods, such as creating fair representations by transforming features to remove sensitive attribute effects, can also be employed. Additionally, reweighting data samples ensures that underrepresented groups receive higher importance during model training.

Model-based strategies

Fair regularization involves adding penalties to the loss function for biased predictions, ensuring metrics such as demographic parity or equal opportunity are met. Adversarial debiasing trains a model alongside an adversarial network, which learns to predict sensitive attributes from the model's outputs, thus reducing bias. Constrained optimization imposes fairness constraints during model optimization, while fair representation learning ensures that latent features are invariant to sensitive attributes while maintaining predictive power.

Post-processing techniques

Threshold adjustment involves modifying the decision threshold to balance performance metrics such as precision and recall. For instance, in the case of class 1 (malignant lesions), increasing the threshold means that the model becomes stricter in classifying an image as malignant. This adjustment can enhance precision—reducing false positives—but often comes at the cost of reduced recall, meaning some malignant cases may be missed.

While these strategies can significantly reduce bias, they often involve trade-offs, such as reduced overall model accuracy. Balancing fairness with performance is a critical consideration.

Retrain With Cleaned-up Data

Clean Data / remove Bias

I chose to remove features with minimal or negative importance to the model, specifically 'anatom_site_general_lower extremity', 'anatom_site_general_posterior torso', and 'anatom_site_general_upper extremity'. This decision simplifies the model by reducing irrelevant inputs, enhancing its interpretability, and minimizing the risk of overfitting. Additionally, I incorporated pos_weight into the BCEWithLogitsLoss function to address the class imbalance in the dataset, ensuring that the model places greater emphasis on correctly classifying malignant skin lesions, which are underrepresented. This adjustment is intended to improve metrics such as recall and F1-score for the minority class without significantly compromising overall model performance.

Model Performance

Week 10

	validation	test
recall	0.64	0.61
pAUC-aboveTPR	0.1034	0.136

Week 11

Validation performance metrics

```
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.98      0.93      0.96      1431
     Class 1       0.25      0.56      0.35        59

    accuracy                           0.92      1490
   macro avg       0.62      0.75      0.65      1490
weighted avg       0.95      0.92      0.93      1490
```

Test performance metrics

```
The partial AUROC of the final model on the test images is 0.1272880171505051
              precision    recall  f1-score   support

     Class 0       0.98      0.95      0.96      1431
     Class 1       0.30      0.58      0.40        59

    accuracy                           0.93      1490
   macro avg       0.64      0.76      0.68      1490
weighted avg       0.95      0.93      0.94      1490
```

	validation	test
recall	0.56	0.58

pAUC-aboveTPR	0.1126	0.127

Recall:

- Validation Recall dropped significantly (**-0.08**), suggesting the model's ability to correctly identify true positives in validation data decreased.
- Test Recall also showed a slight decline (**-0.03**), indicating a reduced performance in generalizing to unseen data.

pAUC-aboveTPR:

- Validation pAUC-aboveTPR increased (**+0.0092**), suggesting an improvement in partial AUC for the validation set, likely at specific TPR thresholds.
- Test pAUC-aboveTPR decreased slightly (**-0.009**), implying a marginal drop in performance on test data.

F1 Score for Class 1:

- **Despite declines in recall**, comparing the F1 scores between Week 10 and Week 11 reveals an **improvement for Class 1** from 0.35 to 0.40, indicating better balance between precision and recall for this specific class. This suggests the model has improved in correctly classifying and balancing errors for Class 1 predictions.

Risks

Medical Misdiagnosis

Incorrect predictions could lead to false reassurance (false negatives), delaying critical medical intervention for malignant lesions, or unnecessary stress and procedures for benign lesions (false positives). Misdiagnosis could damage trust in AI-powered diagnostic tools among patients and clinicians.

Bias in Predictions

The model might perform differently across demographic groups (e.g., skin tones, ages, or genders) due to imbalanced training data, potentially disadvantageous in certain populations. For instance, if most of the training images represent lighter skin tones, the model may be less accurate in identifying lesions on darker skin tones. This could result in higher misclassification rates for individuals with darker skin, exacerbating health disparities and undermining trust in the model among affected groups. Addressing this bias requires curating a diverse dataset, ensuring equitable performance across all demographics.

Privacy Concerns

Handling sensitive patient data (like images and metadata) requires stringent compliance with data privacy regulations (e.g., HIPAA, GDPR). Breaches could lead to legal and reputational issues.