Model Saving

I have saved the model during training on week 9 with torch.save, it was stored at model directory

Dependencies

I have created a text file to document the dependencies in the week 12 notebook.

Deploy method

The best way to deploy the skin lesion classification model depends on the problem statement and requirements. In this case, real-time deployment is the most suitable approach. This strategy is ideal for clinical settings where instant predictions are needed, such as during a dermatologist's consultation. It is also beneficial for scenarios where users upload images and metadata through a web or mobile application and expect immediate feedback. Additionally, real-time deployment supports on-the-fly decision-making, such as prioritizing patients based on the malignancy probability of their lesions. Real-time inference is well-suited to the nature of skin lesion classification, which typically involves small-scale inputs, such as a single image and metadata per request. This approach ensures that users, including doctors and patients, can act promptly on the model's predictions, improving the overall user experience by providing immediate and actionable insights about the lesion's likelihood of being malignant. The other reason that I choose real-time inference is users may expect immediate results, making batch processing less practical.

Performance Metrics to Track in the Monitoring Plan

To ensure the effectiveness and reliability of the skin lesion classification model, several key metrics will be monitored. Model metrics such as accuracy, precision, recall, and F1-score will track overall and class-wise performance, ensuring the model can effectively differentiate between malignant and benign skin lesions. Additionally, AUROC and partial AUC will measure the model's ranking ability, particularly its sensitivity to detect malignant cases. Log loss will be used to evaluate the calibration of predicted probabilities, while prediction latency will track the time required to process inputs and generate outputs, ensuring responsiveness in real-time applications.

From a business perspective, minimizing the false negative rate for malignant lesions is crucial to avoid missed diagnoses, while controlling the false positive rate ensures patients are not unnecessarily alarmed or referred for additional tests. Monitoring resource utilization will help maintain a balanced operational load in clinical workflows, avoiding bottlenecks caused by excessive referrals or delays.

Finally, operational metrics such as data drift and concept drift will monitor changes in input data distribution or target concepts over time, ensuring the model remains accurate and relevant. Tracking model inference volume will detect unusual activity, such as unexpected spikes or dips in predictions, which could indicate system misuse or technical issues. These metrics collectively ensure that the model remains effective, reliable, and aligned with both clinical and operational needs.

Metric Thresholds

Accuracy

- Green (≥ 90%): Indicates the model is performing exceptionally well overall, correctly classifying a high percentage of all cases. No immediate action is needed.
- Yellow (85-90%): The model's performance is slightly degraded, but still acceptable. This could indicate early signs of potential issues, such as data drift or insufficient generalization.
- Red (< 85%): Signals a significant drop in performance. The model may be making too many errors and is not reliable for production use. Requires immediate investigation and possibly retraining.

Precision (Class 1)

- Green (≥ 80%): High precision ensures most of the cases flagged as malignant are actually malignant, reducing false alarms.
- Yellow (70-80%): Indicates a moderate increase in false positives. Patients might be unnecessarily alarmed, but the model is still functional.
- Red (< 70%): A low precision rate means the model is generating too many false positives, leading to wasted resources and patient anxiety. Requires intervention.

Recall (Class 1)

- Green (≥ 85%): High recall ensures that most malignant cases are detected, minimizing the risk of missing critical diagnoses.
- Yellow (75-85%): A noticeable drop in recall indicates some malignant cases are being missed, which could compromise patient outcomes.
- Red (< 75%): Low recall means a high number of malignant cases are going undetected, posing a serious risk to patients. Immediate action is needed.

AUROC

- Green (≥ 0.9): Indicates excellent discriminative ability. The model reliably ranks malignant cases higher than benign ones.
- Yellow (0.85-0.9): Discriminative performance is slightly reduced, but the model is still functional. This may warrant closer monitoring.

- Red (< 0.85): A significant drop in AUROC indicates the model struggles to differentiate between classes, requiring retraining or redesign.

Log Loss

Log Loss quantifies the uncertainty in the model's predictions by penalizing incorrect or overconfident predictions.

- Green (≤ 0.3): Indicates the model's probability estimates are well-calibrated and reliable.
- Yellow (0.3-0.5): Suggests that the model is making overconfident or uncertain predictions. Calibration or retraining might be needed.
- Red (> 0.5): Poorly calibrated probabilities can lead to unreliable decision-making, making the model unsuitable for production.

False Negative Rate (Class 1)

False Negative Rate (FNR) is the percentage of actual malignant cases (Class 1) that are misclassified as benign.

- Green (≤ 10%): Indicates the model rarely misses malignant cases, ensuring patient safety.
- Yellow (10-15%): A moderate increase in missed malignant cases is concerning and requires investigation.
- Red (> 15%): Missing a high number of malignant cases poses significant risks to patients. The model needs immediate attention and retraining.

Data Drift Score

Data Drift Score measures changes in the input data distribution compared to the training data.

- Green (≤ 0.1): Indicates minimal changes in the input data distribution, and the model can continue functioning as expected.

- Yellow (0.1-0.2): Moderate drift suggests early signs of mismatch between the training and current data. Monitoring is required, and retraining may be considered.
- Red (> 0.2): Significant data drift indicates that the input data has changed drastically, making the model less reliable. Immediate investigation and retraining are essential.

Risk mitigation strategies

Green (No Issues)

- Monitor Continuously: Continue tracking performance metrics such as accuracy, AUROC, and log loss. Regularly compare these metrics against baseline thresholds to detect any potential deviations early.

- Validate Input Data: Conduct periodic data quality checks to ensure the input data aligns with the format and quality of the training dataset. Check for missing, corrupted, or anomalous data that could disrupt predictions.
- Maintain Infrastructure: Ensure the serving infrastructure is stable, with no bottlenecks in latency or throughput. Keep software dependencies up to date and maintain efficient hardware utilization.

- Data Logging: Log predictions, input data, and metadata for audit trails and future analysis. This will also support debugging if issues arise later.

Yellow (Warnings)

- Root Cause Analysis: Investigate potential causes for the drop in performance. For example:
    - Is there seasonal variation in the input data (e.g., different age groups or skin tones during a specific time)?
    - Have there been changes in operational workflows, such as new imaging devices producing slightly different image formats?
    - Is there class imbalance emerging in the recent input data?

- Retrain the Model: Collect a recent batch of labeled data reflecting the current trends and retrain the model. Focus on addressing any drift in the input data or target concepts.

- Apply Lightweight Corrections: Before retraining, consider temporary measures:
    - Reweighting: Adjust the loss function to prioritize underperforming classes.

    - Oversampling: Generate synthetic data or duplicate minority class examples to rebalance the dataset.
    - Threshold Tuning: Adjust decision thresholds to optimize performance metrics for the affected classes.

- Regular Communication: Inform stakeholders of the observed warning signs and the mitigation steps being undertaken to manage potential risks.

Red (Critical Issues)

- Immediate Action: Temporarily pull the model from production to avoid significant misclassification errors. If partial functionality is still viable, restrict the model to flagging only clear cases (e.g., obvious malignant lesions).

- Fallback Solutions: Deploy a simpler, rule-based system or a previously validated model to ensure continued functionality while the main model is retrained or debugged.

- Notify Stakeholders: Immediately alert clinicians, operational teams, and other stakeholders about the risks associated with using the current model. Clearly communicate the expected timeline for resolving the issue.

- Debug and Retrain: Diagnose the issue by analyzing input data, performance metrics, and recent logs. Identify whether data drift, concept drift, or technical errors are the root cause, and retrain the model with corrected and updated datasets.

- Infrastructure Adjustments: If the issue is linked to computational constraints (e.g., latency spikes), reassess resource allocation and infrastructure.

Retraining Frequency

Periodic Retraining (Quarterly)

The model will be retrained with newly labeled data every quarter to incorporate updated trends and ensure robustness.

Event-Driven Retraining

If yellow or red thresholds are consistently exceeded due to data or concept drift, the model will be retrained to restore performance.

Active Learning

Cases with high model uncertainty will be actively labeled and added to the training dataset for better coverage of edge cases.

Data Drift and Concept Drift

Data drift and concept drift are critical considerations for maintaining the performance of a skin lesion classification model. Data drift refers to changes in the input data distribution over time. For instance, the demographics of users, such as age groups or gender ratios, might shift, or new imaging technologies could produce images with different characteristics. If the input features, like image properties or metadata, deviate significantly from the training data, the model's performance could degrade. To mitigate data drift, it is important to regularly monitor the distribution of input features and compare them to the training data. Retraining the model periodically with updated datasets can help it adapt to new trends. Additionally, feature scaling and preprocessing techniques can handle minor shifts in data.

Concept drift, on the other hand, occurs when the relationship between input features and the target variable changes over time. This could happen due to advances in medical understanding, such as new criteria for determining malignancy, or changes in labeling protocols for datasets. Concept drift can cause the model to become less accurate because the relationships it learned during training may no longer apply. To address concept drift, the model's performance metrics, such as precision, recall, and AUC, should be monitored regularly using recent data. Retraining or fine-tuning the model with

updated labels and data that reflect new knowledge is essential. Collaboration with domain experts, such as dermatologists, to periodically validate the model's predictions can also ensure that it remains relevant and accurate in clinical applications.