TF Lite Breast Cancer Detection Week 12: Save and Package Model for Deployment

1. Environment Dependencies

- **Operating System (OS)**
  - Ubuntu 22.04.3 LTS
- **Python Version**
  - Python 3.10.12
- **Python Packages and Versions**
  - All of the required Python packages and their specific versions are listed in the requirements.txt file located in the main branch of our repository.
- **Model Saving**
  - This model has been saved for deployment as a pickle file (.pkl), which ensures compatibility with the documented dependencies.

2. Model Deployment: Batch Mode or Real-Time Inference

For our breast cancer detection model, real-time inference would be our preferred approach to deployment. As our ultimate goal is to deploy a TF Lite application that can aid healthcare providers in making quick and streamlined decisions, real-time inference would be more appropriate. Real-time inference allows for immediate predictions, making it particularly valuable in clinical applications where timely decision-making is crucial, and, in some cases, a matter of life or death. By deploying this algorithm in real-time, we envision a future where healthcare providers can analyze mammography images and receive results during patient appointments— thus streamlining interventions and reducing patient anxiety.

Even though batch mode is efficient for processing large datasets, it is not as suitable for our context because it will inherently cause delays between obtaining medical images and forming a diagnosis. This delay is especially counterproductive in scenarios that require rapid decisions to be made. With these considerations in mind, real-time inference aligns better with our ultimate goal of improving patient care through providing quick and actionable insights during the diagnostic process.

3. Model Monitoring Plan

3.1 Model Performance Metrics

For our TFLite breast cancer detection app, we will be tracking the following metrics:

1. **Precision and Recall**
    a. Utilizing precision and recall in our metrics will ensure that the model minimizes the risk of false positives and false negatives. Where high precision is crucial in avoiding unnecessary stress for patients, high recall is essential when identifying all potential malignancies.
2. **F-1 Score**
    a. The F-1 score is the harmonic mean between precision and recall. This provides a comprehensive metric for evaluating the overall quality of predictions in a sensitive medical context.
3. **Precision-Recall AUC**
    a. This is an important metric to track, especially with our imbalanced dataset where metrics like accuracy are misleading. Since precision-recall AUC evaluates performance across multiple thresholds, it provides a more holistic overview. It is a great pair to the F-1 score.
4. **Inference Time**
    a. This metric measures the time it takes for the model to process and predict on new data. To ensure timely diagnosis, efficient and quick predictions are crucial.
5. **False Negative Rate (FNR)**
    a. FNR is particularly focused on missed detections of breast cancer cases—delaying diagnosis and treatment.
6. **Model Drift**
    a. This will be discussed more in depth in the following sections, but continuously monitoring current performance with historical performance is crucial to detect changes in input data or model behavior.

Enforcing a model monitoring plan with these metrics are key to ensuring the reliability and usability of our model and mitigating the risk of model degradation. This is especially important when dealing with sensitive contexts like medical diagnostics. By continuously monitoring these metrics, we can ensure the model continues to perform at a high standard—thereby minimizing the risks to patients and preserving clinical trust.

3.2 Model Thresholds

The following list shows the chosen threshold levels for each metric we have chosen.

- **Precision**
    - Green: ≥ 95%
    - Yellow: 90%-94%
    - Red: < 90%

- **Recall**
    - Green: ≥ 95%
    - Yellow: 90%-94%
    - Red: < 90%
- **F1 Score**
    - Green: ≥ 0.95
    - Yellow: 0.90-0.94
    - Red: < 0.90
- **Precision-Recall AUC**
    - Green: ≥ 0.95
    - Yellow: 0.90–0.94
    - Red: < 0.90
- **Inference Time**
    - Green: ≤ 200 ms
    - Yellow: 201-500 ms
    - Red: > 500 ms
- **False Negative Rate (FNR)**
    - Green: ≤ 5%
    - Yellow: 6%-10%
    - Red: > 10%
- **Model Drift**
    - Green: No largely noticeable drift detected
    - Yellow: Moderate drift observed in specific subgroups
    - Red: Significant drift across critical variables

We have chosen these model thresholds carefully through researching the industry standard for models with similar tasks. These thresholds set high standards for model performance, which is paramount in breast cancer detection. Our EfficientNet model has the capability to meet and exceed these thresholds, which make these benchmarks reasonable and achievable.

3.3 Risk Mitigation for Green, Yellow, and Red Flags

Based on green, yellow, and red thresholds, we plan on taking these risk mitigation strategies.

1. **Green Flags**
    a. **Action:** Continue to consistently monitor the model. To ensure model metrics are stable, continue to perform periodic checks. We will validate the model with subsets of new data at regular intervals.
    b. **Mitigation:** At this threshold, there is no immediate action needed but continue to be vigilant for potential issues that arise.
2. **Yellow Flags**
    a. **Action:** Investigate any root causes of the decline in performance.
        i. Observe any changes in data distribution
        ii. Re-examine the preprocessing function and input data quality
        iii. If the issue is model drift, retrain the model on more recent and updated data.
    b. **Mitigation:** Adjust the hyperparameters, pre-processing function, or fine-tune the existing model again. Increase the frequency of data collection and performance evaluations.
3. **Red Flags**
    a. **Action:** Promptly pull the model from production. While investigating, implement a back-up model or manual review system.
    b. **Mitigation:** A comprehensive audit of the pipeline should be enforced. Before redeployment, the model must be retrained using updated and corrected data and undergo performance validation again.

For the frequency of retraining, regular retraining every 6-12 months is reasonable to incorporate new data and account for any subtle shifts in data patterns. In case of an adverse event occurring, retraining is initiated in response to significant data drift, declining performance metrics, or the introduction of new imaging technologies or patient demographics. Taking this approach will ensure that the model stays current with evolving medical practices and continues to deliver effective and consistent performance in real-world scenarios.

4. Impact of Data Drift and Concept Drift on Mammography Implementation

In the context of mammography images, both data drift and concept drifts may potentially jeopardize the model's performance over time.

4.1 Data Drift

Essentially, data drift refers to any changes in the input data distribution. In our case, data drift could occur due to changes in image acquisition technology, patient demographics, or image preprocessing techniques. For instance, new imaging technologies or updates in mammogram machines may generate images with different resolutions, contrasts, or even noise levels compared to the training data. Naturally, this leads to the model struggling with generalization.

4.2 Data Drift Mitigation Strategies

To ensure the model remains effective over time, a potential way to mitigate data drift could include **periodic retraining with new and representative data** to capture changes in image characteristics. Moreover, **online learning algorithms** could potentially help with enhancing adaptability by updating the model in real-time as new data releases. Additionally, **monitoring the importance of specific individual features** and dropping those that lose relevance over time may be another approach. Lastly, **adapting an ensemble model approach** may improve model robustness. Utilizing multiple models with varying assigned weights could potentially address shifting data distributions.

4.3 Concept Drift

Concept drift in mammography can occur when the relationship between the input features and the target variable (e.g: malignancy for our use case) changes. This may be due to changes in the definition of malignancy, evolved understanding of breast cancer subtypes, introducing new diagnostic criteria impacting how pathology is classified, or even variations in radiologist interpretations.

4.4 Concept Drift Mitigation Strategies

To reduce the risk of concept drift, taking an adaptive approach is crucial. Employing techniques like **ADWIN (Adaptive Windowing)** can aid in monitoring and detecting changes in model performance over time. Calibration drift detection methods, such as **CDDM (Calibrated Drift Detection Method)**, may also help with tracking model calibration and initiate updates when necessary. **Trigger-based ensemble methods** may also be an option, as they can provide rapid adaptation to changes in data distributions— warranting continued accuracy. Additionally, more advanced techniques such as **stacking fast Hoeffding drift detection (FHDDMS)** may identify both abrupt and gradual concept drifts while minimizing false negatives, ultimately improving the model's generalizability.

4.5 Overall Mitigation Strategies

To ensure that the breast cancer detection algorithm generates reliable and consistent performance, implementing a comprehensive strategy is crucial. Below are the following recommendations for implementation:

1. **Regularly Monitoring:** By regularly assessing the model's performance against a curated "golden" dataset, we can accurately and pro-actively identify performance degradation.
2. **Human-in-the-loop Approach:** Consulting expert radiologists to validate and refine predictions will be even more important when potential drift is detected.
3. **Multi-center Validation:** This ensures that the model is tested on diverse datasets from multiple hospitals and promotes robustness by considering variability in image acquisition and patient demographics.
4. **Version Control:** Maintaining meticulous version control such as tracking model versions, training data, and performance metrics, fosters trust and accountability.
5. **Regulatory Compliance:** Lastly, it is crucial to ensure that all updates or adaptations to the model adhere to applicable medical, state, and federal regulations. This includes safeguarding patient data, promising ethical usage, and maintaining compliance with regulatory standards.

All of these strategies work in conjunction with each other to create a strong system that still maintains performance even after advancements in imaging technology, changes in clinical understanding, and changes in patient demographics.