Topic: Analyzing Risk, Bias, and Ethical Concerns Alice Tang and Yinda Chen

TF Lite Breast Cancer Detection Week 11: Analyzing Risk and Bias in Model and Ethical Concerns

1. Identification of Important Features in Model

In order to determine feature importance, we ran a Random Forest model. The following features were identified as top predictors of pathology for breast cancer classification in mammography images:

1. **Assessment**
    a. Importance: Highest (0.38)
    b. Clinical Relevance: Reflects the evaluation of abnormalities by radiologists, which directly influences the likelihood of malignancy.
2. **Calcification Distribution**
    a. Importance: High (0.22)
    b. Clinical Relevance: Describes the spatial arrangement of calcifications, which is often associated with specific types of pathology.
3. **Calcification Type**
    a. Importance: High (0.17)
    b. Clinical Relevance: Represents the nature of calcification, which is essential when determining benign or malignant cases.
4. **Subtlety**
    a. Importance: Moderate (0.08)
    b. Clinical Relevance: This variable indicates the difficulty in visually identifying abnormalities, which may affect how certain a diagnosis may be.
5. **Breast Density**
    a. Importance: Moderate (0.06)
    b. Clinical Relevance: Having higher density can potentially obscure abnormalities, therefore complicating detection and increasing the risk of cancer.

Other predictors include abnormality ID, left or right breast, and image view, which obtained a smaller feature importance score but are still factors that contribute to nuanced distinctions in classification.

1.2 Predictions for Selected Samples

We evaluated the Random Forest model based on five random samples from the test dataset in order to understand the impact of key features on predictions. Below are the results:

- **Sample 1**
    - **Features:** High breast density (3), right breast (1), CC image view (1), calcification type (14), distribution (9), low subtlety (2), and assessment score (1).
    - **True Label:** 1 (benign)
    - **Predicted Label:** 1
    - **Key Factors:** The low assessment score and the distributed calcifications strongly influenced the benign prediction.
- **Sample 2**
    - **Features:** High breast density (3), left breast (0), MLO image view (0), calcification type (11), clustered distribution (4), subtlety (3), and assessment score (4).
    - **True Label:** 2 (malignant)
    - **Predicted Label:** 2
    - **Key Factors:** The high assessment score and clustered calcifications were decisive in predicting malignancy.
- **Sample 3**
    - **Features:** High breast density (3), left breast (0), CC image view (1), calcification type (18), diffuse distribution (0), low subtlety (0), and assessment score (3).
    - **True Label:** 0 (normal)
    - **Predicted Label:** 0
    - **Key Factors:** The widespread distribution and low subtlety aligned with a normal classification.
- **Sample 4**
    - **Features:** Low breast density (1), left breast (0), MLO image view (0), calcification type (41), clustered distribution (4), subtlety (3), and assessment score (1).
    - **True Label:** 1 (benign)
    - **Predicted Label:** 1
    - **Key Factors:** The combination of clustered calcifications and low assessment score contributed toward a benign diagnosis.
- **Sample 5**
    - **Features:** Very low breast density (0), right breast (1), MLO image view (0), calcification type (18), diffuse distribution (0), high subtlety (4), and assessment score (4).
    - **True Label:** 2 (malignant)
    - **Predicted Label:** 2

- **Key Factors:** The high assessment score and calcification type were essential in determining malignancy.

1.3 Feature Changes to Impact Predictions

In order to alter the model's predictions, these are the following features that were identified as highly influential from our analysis:

- **Assessment Score:** Even a small increase in this score may significantly impact predictions toward malignancy.
- **Calcification Distribution:** Moving from a widespread distribution to a clustered pattern could increase the probability of malignancy.
- **Calcification Type:** Changes to a type that is more associated with malignancy (for example, from 11 to 41) may also impact predictions.
- **Subtlety:** Lowering subtlety scores could flip predictions from benign to normal— thus making for easier detection.
- **Breast Density:** Lowering the density may potentially improve the detection of abnormalities, potentially impacting predictions.

Essentially, this analysis underlines the importance of clinical features in contributing to the model's predictions and highlights potential areas for any further needed improvements to enhance interpretability and performance.

2. Protected Categories in Data

Our dataset technically included protected categories such as PatientBirthDate and Gender in the dicom_csv file. However upon further inspection, every entry was completely empty, thus we were unable to use it in our analysis. Additionally, our model does not directly use these categories, as it primarily focuses on the image data to make predictions. Instead, we explored variables such as breast density, which may be indirectly correlated with protected attributes— for example, breast density often decreases in age. This ensures that our model utilizes meaningful predictors while avoiding any direct reliance on incomplete data fields.

3. Bias of Model

Based on the data, the correlation matrix and breast density distribution offer insight into potential biases in the model. While our dataset includes empty fields for protected categories such as PatientBirthDate and Gender, these are not directly used in the model since it focuses on image data. However, we did decide to examine breast density as a potential variable that may be indirectly correlated with protected attributes like age, since breast density typically decreases with age.

From the correlation matrix that can be found in this week's notebook, breast density does show a weak positive correlation with pathology_numeric (0.0766) and assessment (0.1773), but a weak negative correlation with subtlety (-0.2161). Essentially, this implies that while breast density is not strongly correlated with pathology or assessment outcomes, it may still impact predictions in subtle ways.

The breast density distribution across pathology classes does show some potential patterns. For example, higher breast density values (1 and 2) are more common in benign categories, whereas lower breast density values (3 and 4) are more evenly distributed across benign and malignant categories. This shows that the model may inherently place more weight on specific density categories, which potentially influences its predictions.

To address these potential biases, we must monitor how the model's performance varies across the different breast density groups. If significant differences are seen, we may need to incorporate techniques like rebalancing or adding certain fairness constraints in order to ensure the model treats all of the categories fairly.

4. Bias Removal Strategies and Their Impact on Predictions

Based on our analysis of bias in breast density, these are a few strategies to mitigate bias, along with any potential impact on model performance and fairness:

Strategy 1: Data Balancing

This strategy focuses on ensuring that there is a balanced representation of pathology outcomes across each of the breast density categories in order to mitigate biases in the model's predictions. This would involve two main techniques: oversampling, where samples from underrepresented groups are replicated (e.g: benign cases in dense breasts) and undersampling, where samples from overrepresented groups are reduced (e.g: malignant cases in high-density breasts). We would then need to use these adjusted datasets for model training.

There are multiple pros and cons with this strategy. On one hand, it may reduce bias in predictions since it prevents the model from disproportionately associating high breast density with malignancy and improves generalization for underrepresented groups. However, it may also introduce the risk of overfitting when the minority class is oversampled excessively.

Strategy 2: Feature Removal or Transformation

This strategy involves either moving "Breast Density" as a feature during training, or potentially transforming it to a less sensitive representation to minimize influence on the model. Alternatively, we could encode breast density as a binary feature, like high density or low density, which helps with

simplifying representation. This would include excluding breast density from the input features or replacing it with a more general representation to reduce its potential impact.

Benefits of this approach include eliminating any sort of direct reliance on a potentially sensitive feature, and would encourage the model to prioritize clinically relevant features, like calcification type or distribution. The drawback would be that there would likely be a decrease in predictive accuracy and F-1 score, as breast density most likely holds clinical value in many cases.

Strategy 3: Reweighting Samples

This strategy aims to address class imbalance through assigning higher weights to samples for the minority group, to ensure that all groups are contributing equally to the goal. This specific implementation would include calculating sample weights inversely proportional to the frequency of each breast density-pathology combination and passing these weights to the model during training.

The advantages of this approach are clear— it will effectively address class imbalance without needing to oversample, and still maintains the diversity of the dataset. However, there will be an added layer of computational complexity during the training process, which may be a barrier.

5. Potential Risks of Model on Stakeholders

When deploying a model for detecting breast cancer, there are several risks to consider when regarding its impact on various stakeholders.

1. **Patients (Citizens/Customers):**
    a. **False Negatives:** If the model is unable to identify malignant cases correctly, it may delay critical diagnoses and treatments, severely affecting patient outcomes.
    b. **False Positives:** If the model overpredicts malignant cases, this could lead to psychological damage to patients, as it causes unnecessary stress and may lead to unneeded medical tests and increased healthcare costs.
2. **Healthcare Providers**
    a. **Overly Relying on AI:** If providers become overly dependent on the model's predictions, this could in turn lead to undermining the importance of having human expertise and more nuanced and tailored clinical decision-making.
    b. **Liability Risks:** Any errors in the model predictions could lead to misdiagnosis, which would then raise questions about accountability and liability.
3. **Government**

   a. **Compliance:** Deploying this model must meet regulatory requirements for ethical, equitable, and accurate use. Without addressing potential biases, this could lead to non-compliance.
   b. **Data Privacy:** Naturally, data privacy is paramount when handling sensitive patient data. If there is misuse or breach, this could snowball into legal consequences of the public distrust surrounding AI/ML.

4. **Environment**
   a. **Resource Consumption:** Deploying and training ML models may lead to contributing heavily to energy use and environmental impact.
5. **Healthcare Equity**
   a. **Contributing to Present Disparities:** If the model only considers data from certain demographic groups or regions, it may not generalize well to underrepresented communities— further contributing to healthcare inequalities.

This week's assignment is incredibly important in considering risks and potential biases in our model. To mitigate these risks, we must ensure transparency, conduct a thorough bias and fairness analysis, and continue to validate the model in real-world scenarios. When deploying this model, collaborating with healthcare professionals, policymakers, as well as patient advocacy groups may also help to address stakeholder concerns effectively.