Topic: Engineer Features Alice Tang and Yinda Chen

TF Lite Breast Cancer Detection Week 5: Feature Engineering

1. Introduction

This document outlines the data processing and augmentation techniques applied to the "calc_case" and "mass_case" training sets, focusing on enhancing their utility for ML models. The preprocessing steps included label encoding, feature selection using the Fisher Score, feature transformation, and dimensionality reduction. These techniques helped refine the datasets by keeping only the most influential variables and eliminating irrelevant ones, improving model efficiency and reducing overfitting risks. Additionally, data augmentation was applied to mammogram images through random rotations, scaling, and Gaussian blur to increase data diversity and robustness, ensuring the model can better generalize to real-world clinical scenarios. We aim to take a holistic approach to optimizing both dataset quality and model performance.

2. Data Processing for "calc_case" Training Set

Before the calc_case training set was ready for modeling, we had to conduct several essential preprocessing steps. Virtually, these steps focused on feature selection, feature transformation, and dimensionality reduction, all part of enhancing the dataset's utility for future ML models. Through carefully refining the dataset, we retained only the most significant features, while redundant or irrelevant ones were systematically eliminated to improve model performance and efficiency.

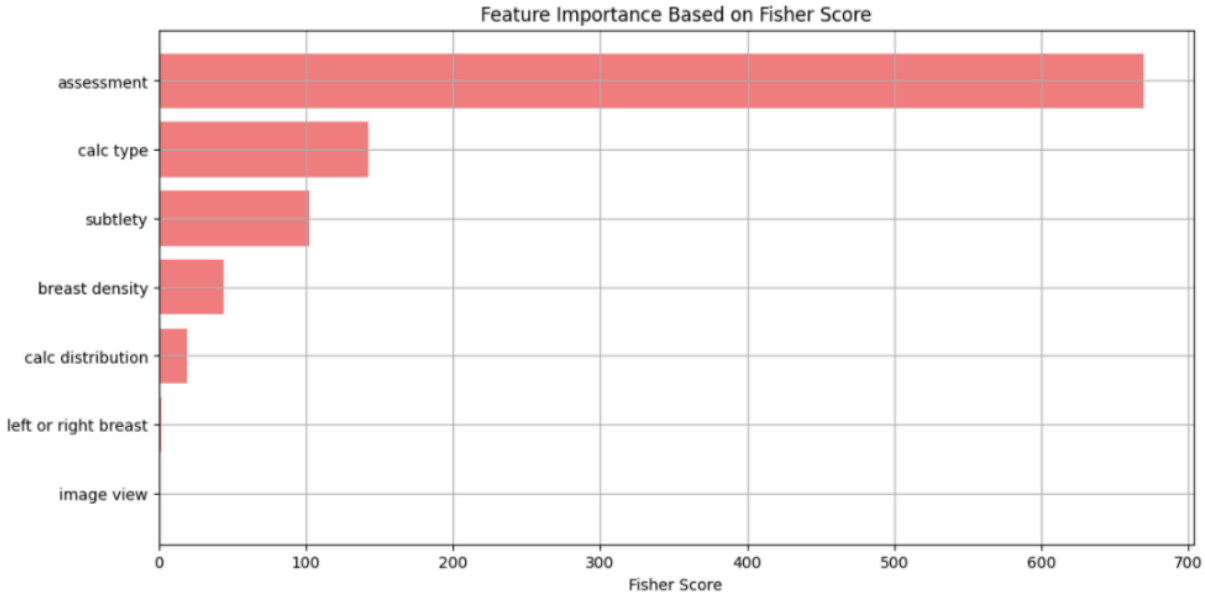
2.1 Label Encoding of Categorical Variables

We first started with encoding categorical variables, which we had done previously in last week's cleaning. However, we focused on transforming the pathology column specifically, which indicates whether a case is benign or malignant, needed to be converted from categorical to numerical format. This was a crucial step, as most ML algorithms are unable to process categorical data directly. Through assigning a unique numerical label to each class (benign or malignant), we made the target variable interpretable by the classification model, thus facilitating an accurate prediction process.

2.2 Fisher Score for Feature Selection

To determine which features contributed the most to distinguishing between benign and malignant cases, we utilized the Fisher Score method. Essentially, the Fisher Score is a statistical feature selection technique that calculates the ratio of between-class variance to within-class variance for each feature. Achieving a higher Fisher Score implies that the feature is highly discriminative between classes.

By applying this method, we found key features such as assessment, calc type, and subtlety were found to be crucial in classifying cases. On the other hand, features such as left or right breast and image view, which obtained lower Fisher Scores, were deemed less informative. Our aim was to reduce the dimensionality of the dataset, since this would allow the model to focus on features that had a larger impact on the classification task. With this in mind, we removed these less important features. This reduction also helps in mitigating the risk of overfitting, where the model may potentially perform well on training data but poorly on unseen data due to excessive complexity.



2.3 Feature Combination and Transformation

In certain cases, we combined specific features to further streamline the dataset while preserving its most valuable information. For instance, the calc type and calc distribution columns were merged into a single feature, calc_type_distribution. These two features were closely related and combining them not only reduced dimensionality, but also maintained the interpretative value of the data. This feature engineering was very intentional, as it aids in simplifying the dataset without sacrificing predictive power— ultimately improving the model's generalization.

2.4 Dimensionality Reduction

Following the Fisher Score analysis, we removed additional columns to further reduce dimensionality. In this process, we eliminated not only less significant features but also irrelevant data like patient IDs, image file paths, and other non-predictive variables. These columns would only add to the noise in the dataset and would not contribute to the classification task at hand.

By performing this dimensionality reduction, we reduced the number of input variables, making the dataset more manageable and allowing machine learning models to process data more efficiently. In turn, this improves the training speed for models and reduces computational costs while keeping the quality and relevance of the data.

2.5 Final Dataset Preparation

After applying feature engineering and dimensionality reduction, the final version of the calc_case training set was ready for modeling. This version is comprised of only the most important features, ensuring the categorical variables were properly encoded and irrelevant data was removed. In the end, the dataset was ready for future modeling, allowing classifiers to focus on the most predictive variables, thereby elevating both the model's accuracy and its efficiency during training.

3. Data Processing for "mass_case" Training Set

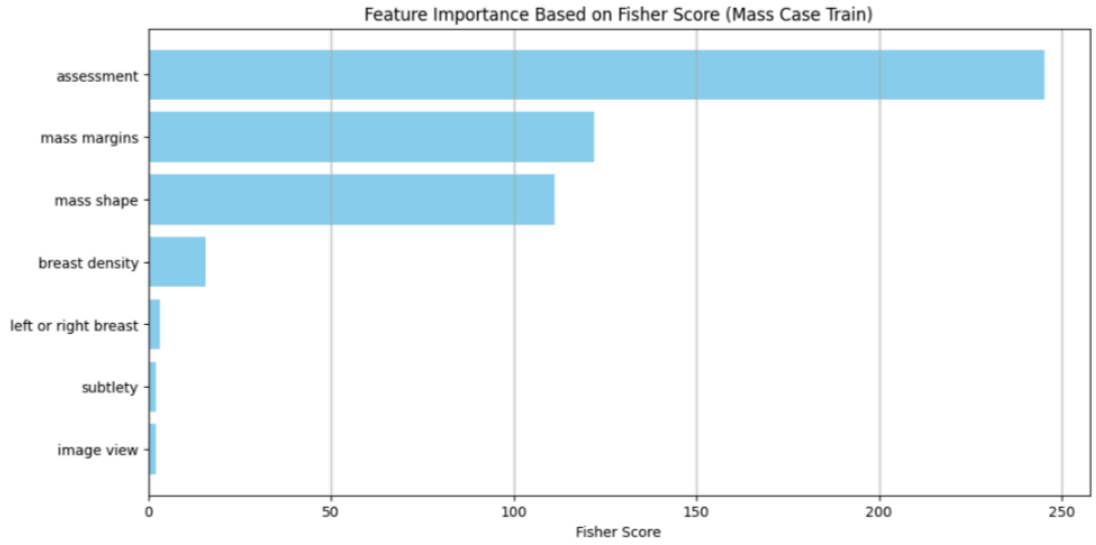
Similarly, in the mass_case training set, we followed the same series of preprocessing steps to improve the dataset's structure for ML models. Once again, the overarching goal was to preserve the most informative variables while eliminating irrelevant ones. This structured approach was crucial for enhancing model performance and ensuring robust classification outcomes.

3.1 Label Encoding for Target Variable

The next crucial preprocessing step was the encoding of the target variable, pathology, which classifies each mass as benign or malignant. To make this variable compatible with ML models, it was transformed from categorical format to numerical format using Label Encoding. This conversion is essential because ML algorithms handle numerical values more efficiently than categorical data. By assigning distinct numerical labels to each category, the encoding simplifies the classification task, allowing the model to accurately recognize and learn from the target variable effectively.

3.2 Feature Selection Using Fisher Score

Once again, to identify the most relevant predictor variables for classification, the Fisher Score method was employed.



For the mass_case dataset, the key features identified through this process included assessment, mass shape, and mass margins, which showed strong discriminatory power. Conversely, features such as left or right breast, subtlety, and image view returned lower Fisher Scores and were deemed less significant. This feature selection process was critical for reducing dimensionality and ensuring that the retained features had the greatest predictive value. By concentrating on the most relevant features, we simplify the model, reducing its complexity and lowering the risk of overfitting.

3.3 Feature Transformation and Dimensionality Reduction

Several transformations were applied to further optimize the dataset:

1. **Conversion of Categorical Variables:** Features like mass shape, mass margins, and pathology were converted to categorical data types. This conversion aids in consistency and ensures efficient processing by ML models, which often require categorical data to be explicitly defined for proper handling.
2. **Elimination of Redundant Features:** Irrelevant features such as patient_id, image paths, and abnormality IDs, which do not contribute to prediction, were systematically removed. Additionally, columns that exhibited no variance, such as abnormality type, were also dropped. This streamlining of the dataset minimizes noise, thus enhancing the model's focus on meaningful features.

These transformations significantly reduced the dataset's dimensionality, enabling a more efficient model training. By removing irrelevant and redundant features, we ensured that the dataset retained only the most informative variables, ultimately leading to improved model performance.

4. Data Augmentation for Mammogram Images

In addition to dimensionality reduction we applied on the other datasets, we also conducted data augmentation on the mammogram images in order to enhance the diversity of the training data. Data augmentation is paramount for improving model generalization, especially in medical image analysis, where datasets are often limited (as in our case). The different augmentation steps we took included image rotation, scaling, and the application of Gaussian blur to simulate various real-world scenarios and improve the robustness of the model.

4.1 Rotation and Scaling

To introduce variability in the dataset, we applied random rotations and scaling transformations to the mammogram images. Each image was rotated by a randomly selected angle within a predefined range and scaled by a factor between 0.8 and 1.2. This technique improves the model's robustness since it becomes invariant to changes in orientation and size, which are common in real clinical environments.

For this, specifically, we used the "rotate_and_scale_image" function, which:

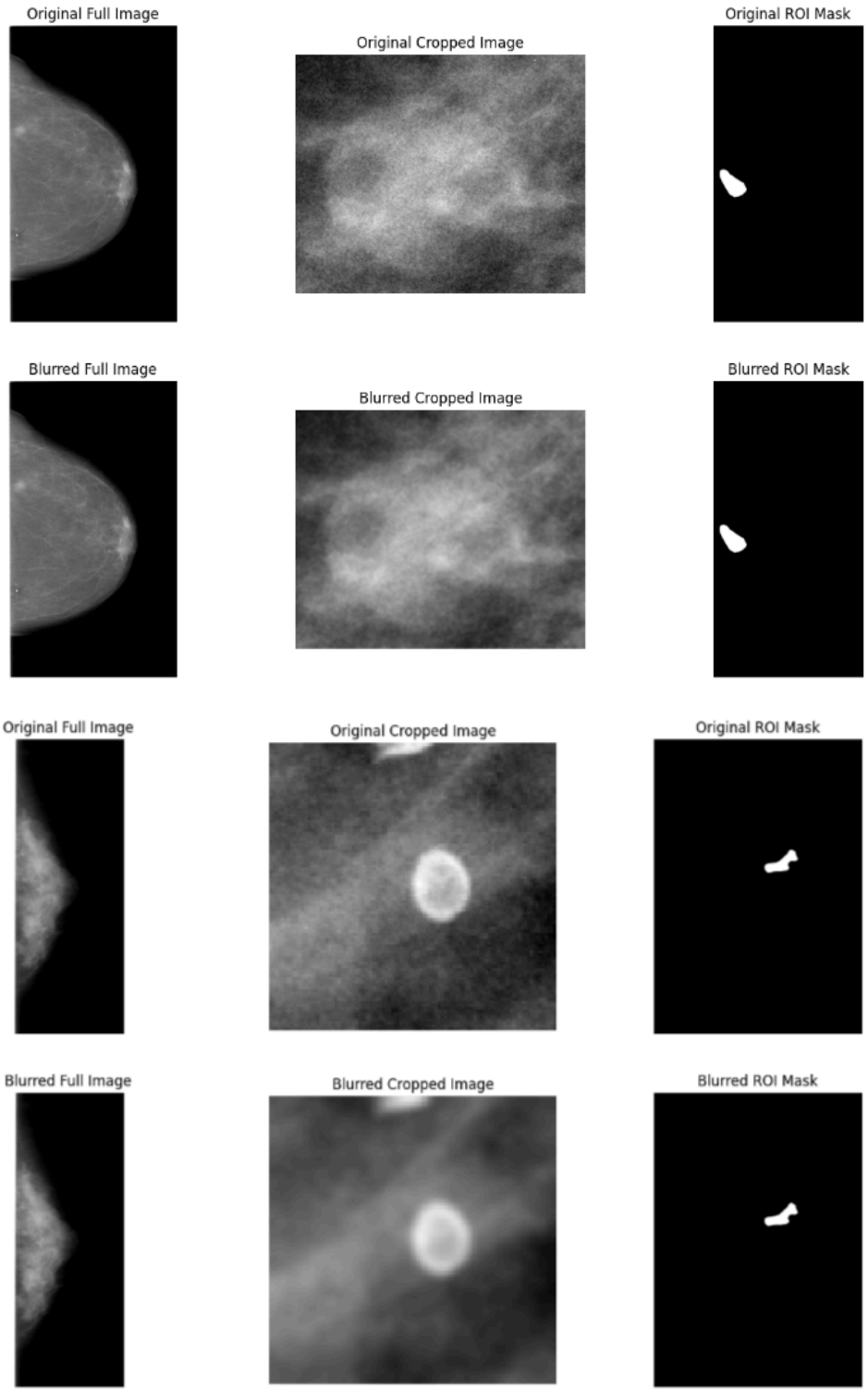
- **Rotates the image** by calculating its center and applying an affine transformation.
- **Scales the image** based on the scale factor— either zooming in or out.
- Ensures the **transformed image is resized to its original dimensions** by either cropping or adding padding where necessary.

This process was applied to the full mammogram images, cropped regions, and the ROI (Region of Interest) mask images, generating augmented versions of each.

4.2 Gaussian Blur

To enhance the dataset further, Gaussian blur was introduced as another augmentation method. This method simulates variations in image quality, such as slight out-of-focus areas, which naturally occurs in medical imaging. The "apply_gaussian_blur" function was used to apply random levels of blur by adjusting the kernel size and the standard deviation of the Gaussian distribution.

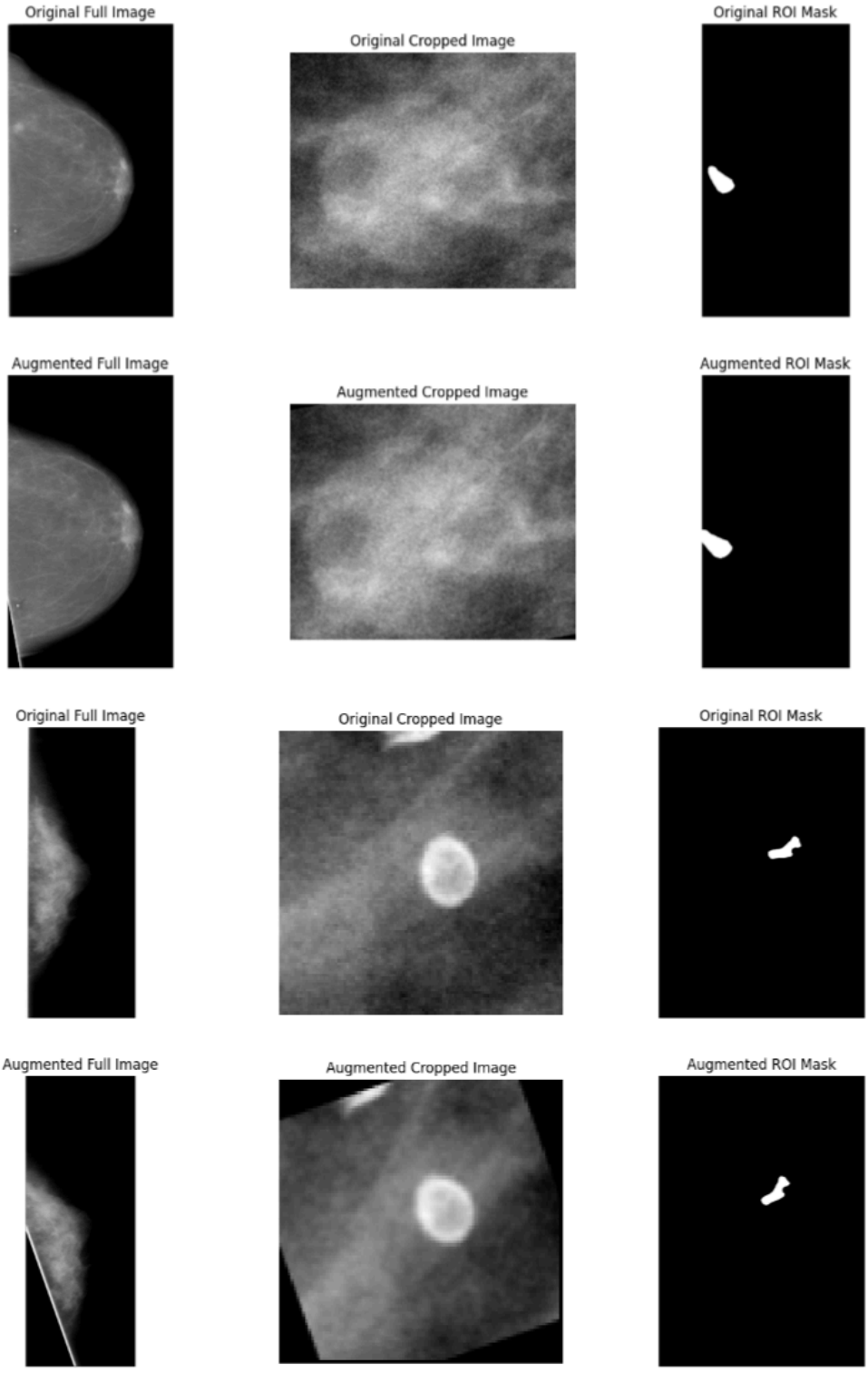
This augmentation was applied across all image types, including full mammogram, cropped, and ROI mask images. By incorporating Gaussian blur, it ensures the model becomes more robust to variations in image sharpness, improving its ability to generalize to real-world cases.



4.3 Visualization

For clarity and comparison, we made sure to visualize each augmented set by setting the original and transformed images side by side. The visualizations offered valuable insights into how the

augmentations impacted the image structure, confirming that the modifications remained realistic and suitable for model training. The modified and original images are shown below.



These augmentation techniques substantially expanded the training dataset, improving the model's ability to detect masses under diverse conditions. As a result, the model’s generalization capabilities and overall performance have the potential to be significantly improved.

5. Conclusion

In summary, the thorough preprocessing and augmentation strategies employed for the calc_case and mass_case training sets significantly enhance the datasets' quality and utilization for ML applications in breast cancer detection. By emphasizing effective feature selection, transformation, and dimensionality reduction, we maintained he most crucial variables while eliminating unnecessary noise and redundancy. Additionally, the incorporation of data augmentation techniques enriched the training data, facilitating an environment for model robustness and generalization. This integrated approach not only enhances predictive model performance but also establishes a strong foundation for future innovations in medical imagery. By prioritizing high-quality data, we aim to enhance the accuracy and reliability of breast cancer detection systems, ultimately supporting better patient outcomes.