Ari Pai (ap66552)
Vishwa Patel (vp8792)
Aryan Shah (ahs2388)
Milan Vaghani (mhv386)
Sankeerth Viswanadhuni (vps386)

Skin Cancer Detection Based on 3D Total Body Photography

Introduction

The aim of this project is to develop a machine learning model capable of detecting malignant skin lesions based on diagnostically labeled images extracted from 3D total body photography (TBP). These images will simulate smartphone-quality photos and include metadata on patient demographics and lesion characteristics. By leveraging strongly labeled histopathology-confirmed cases and weakly labeled benign cases, this system seeks to improve early skin cancer detection, particularly in underserved populations with limited access to dermatological care.

Project Architecture

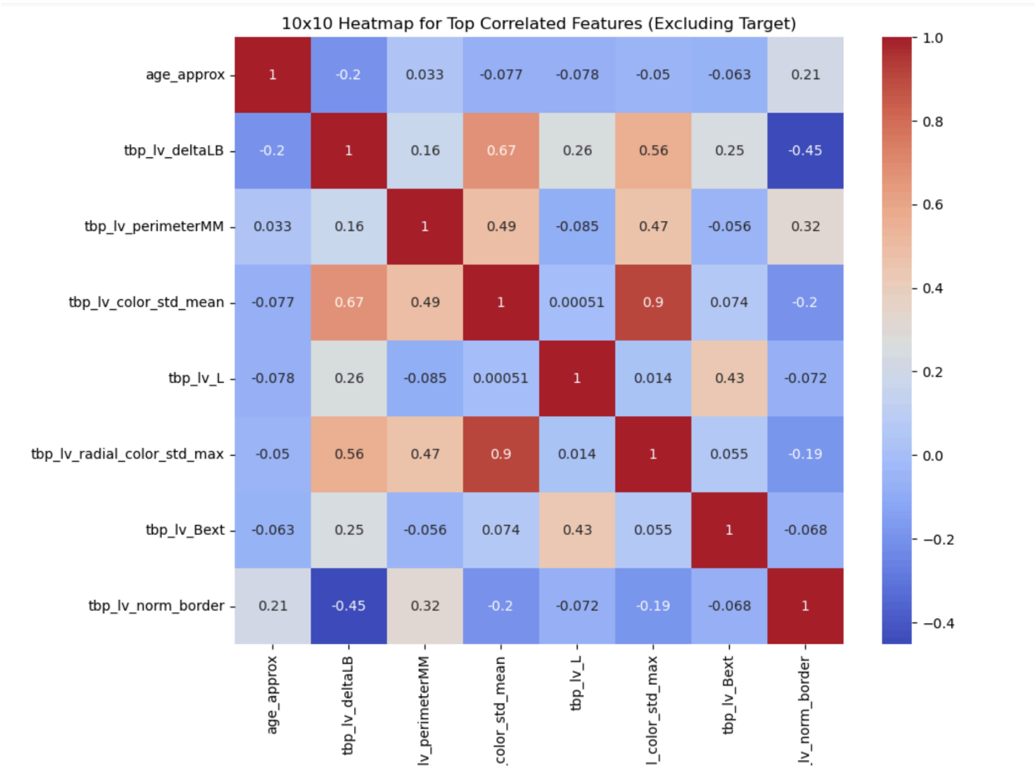
1. **Dataset Collection**: The dataset will consist of cropped images from the SLICE-3D dataset, paired with metadata such as patient demographics and lesion characteristics (e.g., lesion size, location, color irregularity). The training set includes both confirmed malignant lesions and benign cases classified by dermatologists.
    a. Link to Data
2. **Data Preprocessing**: Images will be preprocessed by resizing and normalizing pixel values. The metadata, which includes patient age, lesion size, color contrast, border jaggedness, and anatomical site, will be standardized for easier integration into the model.
3. **Model Selection**: We will experiment with various convolutional neural networks (CNNs) and models that incorporate both image data and metadata, such as ResNet architectures or hybrid models combining CNNs with metadata inputs.
4. **Training the Model**: The model will be trained to learn the relationships between image features (e.g., lesion irregularity) and input metadata (e.g., patient age, anatomical site) and the binary target variable (malignant or benign). We aim to condition the model to differentiate lesions with high sensitivity to maximize clinical usefulness.
5. **Skin Cancer Detection**: The model will predict the probability of malignancy for each lesion image, taking into account both visual features and metadata, and assign a score ranging from 0 to 1, with the positive class being malignancy.
6. **Evaluation and Improvement**: The model will be evaluated based on the partial area under the ROC curve (pAUC) above 80% true positive rate (TPR), as this metric prioritizes sensitivity in identifying malignant cases.

Dataset Collection

The primary dataset for this project will be the SLICE-3D dataset, which includes about 400,000 images representing every lesion from thousands of patients across different continents. These images are standardized 15 x 15 mm lesion crops captured with 3D TBP technology, mimicking non-dermoscopic, smartphone-quality photos. The metadata includes both lesion-specific attributes (e.g., size, color irregularity) and patient demographic information (e.g., age, sex).

Preprocessing the Data

- **Image Preprocessing**: JPEG images will be resized and normalized to standard dimensions suitable for input into the CNN. Augmentation techniques (e.g., rotation, flipping) will be applied to improve generalization.
- **Metadata Standardization**: Key metadata fields, such as lesion size (in mm), anatomical site, and patient age, will be standardized and encoded to ensure consistent interpretation by the model.
- **Dimensionality Reduction:** After performing an initial EDA of the train-metadata.csv we see that there are 2 features which are highly correlated with each other which creates multicollinearity problem which can affect the performance of machine learning model, as it can make it harder to isolate the effect of individual features on the target variable. We might possibly use PCA of T-SNE to reduce the dimensionality



Model Selection

We will explore various deep learning architectures for both image and metadata processing to find the one with the highest accuracy. Our proposed

- **Convolutional Neural Networks (CNNs)**: Create a CNN which will process the lesion images, focusing on visual features such as color irregularity and border jaggedness. We could also use a hybrid approach, which will integrate image features from CNNs with metadata (e.g., age, lesion size) using a separate feedforward neural network, merging both streams before the final classification layer. The final classification will output a probability from 0 to 1, where malignancy is the positive class.
- **Transfer Learning**: We may also consider using pre-trained models like ResNet. We can fine-tune these models on the SLICE-3D dataset to leverage existing knowledge from large image datasets.

Training the Model

- **Loss Function**: Binary cross-entropy will be used to train the model for binary classification. Class weighting may be applied to account for the imbalance between benign and malignant cases.
- **Optimization**: We will use Stochastic Gradient Descent to optimize, and will experiment with enhancements such as SGD with momentum or Adam. We will also consider the use of regularization techniques to prevent overfitting.

Evaluation and Improvement

- **Primary Metric**: The model will be evaluated based on the pAUC above an 80% TPR, focusing on high sensitivity to maximize detection of malignant lesions.
- **Model Tuning**: Hyperparameters such as learning rate, CNN depth, and batch size will be fine-tuned to optimize model performance.

Potential Challenges

- **Imbalance in Data:** A large portion of the image set consists of benign cases (only 393 are malignant). This could be a challenge when training the model and may compel us to use resampling techniques or class weighting, as mentioned above, to overcome this. It may also require techniques like oversampling or SMOTE to balance the dataset.
- **Heterogeneity of Lesions**: Lesions may vary significantly in appearance, making it challenging to build a generalizable model.