Problem Statement

Skin cancer is the most common form of cancer in the United States and ranks 17th globally (WCRF). There are three major types of skin cancer—Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and Melanoma. While BCC and SCC are considered less lethal, melanoma is the deadliest form of skin cancer. It is expected to be diagnosed over 200,000 times in the US in 2024, with nearly 9,000 deaths projected. Automated image analysis tools play a significant role in expediting clinical presentation and diagnosis, positively impacting hundreds of thousands of people each year.

For a telehealth app company, addressing the challenge of skin cancer detection in underserved populations or non-clinical settings is particularly significant. Current diagnostic methods rely on high-quality dermatoscope images, which are typically captured in dermatology clinics. These images reveal morphological features not visible to the naked eye.

To provide this early detection service on our platform, we need to develop an algorithm capable of accurately classifying lower-quality malignant skin lesions from benign ones. Additionally, this algorithm should assist in diagnosing users based on their type of lesions and personal information.

Articulation of Value

Early and accurate detection

Early and accurate detection of skin cancer is essential for effective treatment and improved patient outcomes. By enabling the analysis of lower-quality images from telehealth platforms, this algorithm facilitates earlier diagnosis, especially in underserved populations where access to specialized dermatologic care is limited. This proactive approach can lead to earlier interventions, reducing the severity of the disease and improving survival rates.

Increased Accessibility

Telehealth platforms provide a convenient and accessible means for individuals to monitor their skin health, particularly in areas where dermatologists are not readily available. The algorithm's ability to accurately classify skin lesions from smartphone images makes skin cancer screening more accessible to a broader audience, addressing disparities in healthcare access.

Cost-Effective Care

By integrating automated image analysis into telehealth services, the need for immediate in-person dermatological consultations can be reduced. This not only lowers the cost of healthcare for patients but also alleviates the burden on dermatology clinics, allowing them to focus on more complex cases that require specialist attention.

Reduction in Emotional and Physical Impact

Early detection of skin cancer allows for less aggressive treatment options, which can significantly reduce the emotional and physical impact on patients. Addressing skin cancer early can lead to less invasive treatments, quicker recoveries, and an overall improved quality of life for patients.

Alignment with Healthcare Missions

This project supports the mission of telehealth companies to provide equitable and accessible healthcare solutions. By bridging the gap in specialized care and enhancing the accuracy of skin cancer detection in non-clinical settings, the algorithm contributes to a more inclusive healthcare system, ultimately improving patient outcomes and addressing healthcare disparities.

Potential Economic Value

From a business perspective, the potential economic value of the app consists primarily of two key revenue streams: partnership fees, and market share growth. Given these assumptions, we can evaluate the potential economic value for the company.

Partnerships and Licensing

The company might enter into additional partnerships[1] with healthcare providers due to new features, generating additional revenue streams.

Calculation:

Partnership Revenue = 5 partnership × \$20,000 = \$100,000

[1] Company might be in partnership with a healthcare provider to offer virtual services to app users. Let’s assume each partnership value costs \$20,000 per year and it is based on partnership agreements.

Market Share growth

With the successful implementation of skin cancer detection in the app, the company could gain a significant market share, potentially leading to increased revenue and enhanced brand value. Assuming that this new feature helps the company capture an additional 1% of the market share, it could result in substantial growth.

Calculation:

Estimated Market Share Increase = 1% × $83.5 billion market share = $835,000,000

Total economic value

$100,000 (partnerships) + $835,000,000(market share) = $835,100,000

Project plan

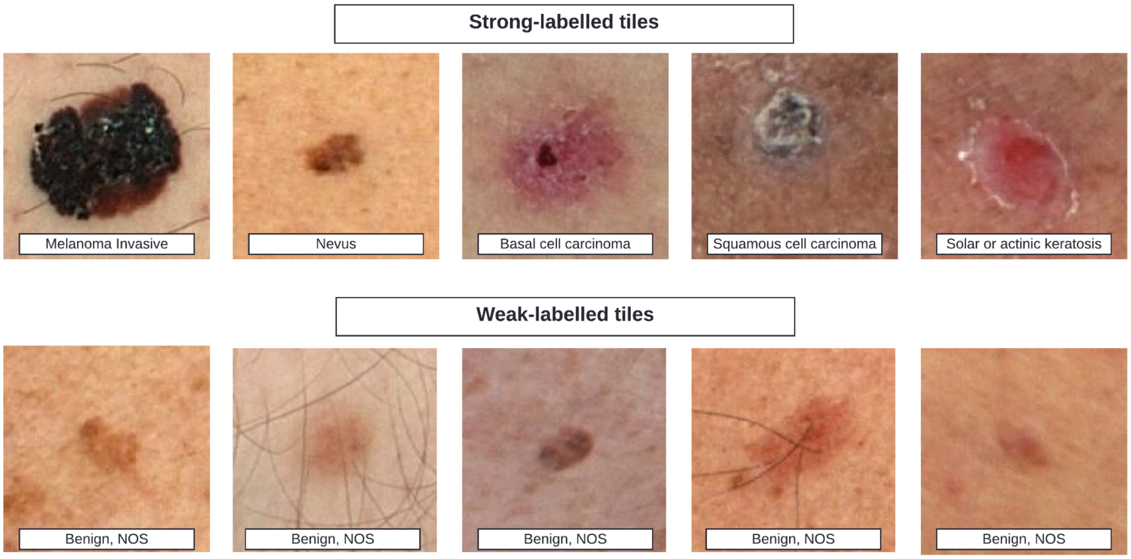
Weeks	Objectives	Steps
Week 1	Identify the problem statement and dataset	● Define problem statement ● Specify goals ● Gather data from reliable source
Week 2	Ingest and explore the dataset	● Load data ● Handling different data format ● Inspect dataset dimensions,types,missing values ● Perform data quality checks
Week 3	Perform exploratory data analysis	● Visualize data with Matplotlib and Seaborn ● Identify anomalies and patterns

		● Assess class distribution and check for imbalance
Week 4	Make data model ready	● Clean data ● Address outliers ● Standardize image pixel values ● Encode categorical data
Week 5	Engineer features	● Feature Extraction ○ Extract relevant features from images ○ Create new features for tabular data ● Feature Selection ○ Use correlation analysis and feature importance
Week 6	Develop 1st modeling approach (simple, the baseline)	● Examples: Logistic Regression for tabular data and a basic Convolutional Neural Network (CNN) for image data. ● Train the model on the training dataset ● Evaluate performance using metrics like accuracy, precision, recall, F1-score.
Week 7	Develop 2nd modeling approach (more complex)	● Advanced CNN architectures (e.g., ResNet, Inception). ● Train the model on the same dataset. ● Use cross-validation and fine-tuning to optimize performance
Week 8	Develop 3rd modeling approach (even more complex)	● Transfer learning with pre-trained

		models, ensemble methods ● Experiment with hyperparameters and additional techniques
Week 9	Select the winning model	● Analyze performance metrics of all models. ● Select and save the best model
Week 10	Data Centric AI	● Set up processes for ongoing evaluation and updates. ● Continuously improve data quality and diversity
Week 11	Explain the model, analyze risk, bias and ethical considerations	● Use tools like SHAP or LIME to explain model predictions. ● Identify and mitigate any biases in the model. ● Ensure the model is used responsibly and respects patient privacy.
Week 12	Save and package your model for deployment. Build your model monitoring plan	● Serialize the model using tools like TensorFlow SavedModel or ONNX. ● Create a deployment package including model files and dependencies ● Set up logging and performance monitoring to track the model's behavior in production.
Week 13	Bring it all together	● Use streamlit or Flask for web app deployment ● Test web app

Dataset

The dataset comprises diagnostically labeled images of skin lesions, provided in JPEG format, accompanied by a .csv file containing metadata. The metadata includes binary diagnostic labels indicating whether each lesion is malignant or benign, as well as additional input variables such as age, sex, and anatomical site. The dataset features standardized cropped lesion images obtained from 3D Total Body Photography (TBP). Each lesion image is a 15x15 mm crop from a high-resolution tomographic image capturing the entire visible skin surface. The images are sourced from thousands of patients seen between 2015 and 2024 across nine institutions and three continents by the International Skin Imaging Collaboration (ISIC) . The institutions involved include Hospital Clínic de Barcelona, Memorial Sloan Kettering Cancer Center, Hospital of Basel, FNQH Cairns, The University of Queensland, Melanoma Institute Australia, Monash University and Alfred Health, University of Athens Medical School, and Medical University of Vienna.The dataset features both "strongly-labeled" tiles, which have been validated through histopathology, and "weak-labeled" tiles, which were deemed benign by a doctor but were not subjected to biopsy. This dataset was found from ISIC 2024 - Skin Cancer Detection with 3D-TBP[2] competition on Kaggle.



[2] ISIC 2024 - Skin Cancer Detection with 3D-TBP (https://www.kaggle.com/competitions/isic-2024-challenge/data)

The dataset addresses the problem statement by providing a comprehensive collection of images and associated metadata for training and evaluating skin cancer detection algorithms. By differentiating between benign and malignant cases based on diagnostic labels, the dataset enables the development of machine learning models that can predict the likelihood of malignancy in skin lesions. The standardized nature of the images and the inclusion of both strongly and weakly labeled cases allow for robust model training and validation. The use of 3D TBP images mimics real-world scenarios where high-quality dermoscopic images might not always be available, thereby improving the app's capability to handle and accurately assess non-dermoscopic images. This helps in early detection and triaging of skin cancer, potentially improving patient outcomes and enhancing diagnostic efficiency in diverse clinical settings.

Modeling

This is a supervised binary classification problem where I will be classifying cases as benign or malignant, with the target variable being 0 for benign and 1 for malignant. My approach is to use a Convolutional Neural Network (CNN) as the base model for image data, then incorporating both metadata and image data to classify the cases. The initial architecture of the model:

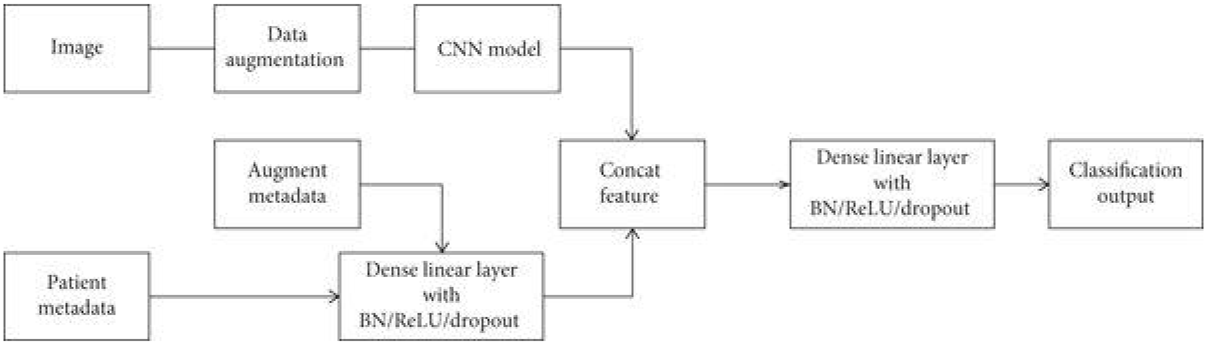


Image
Data augmentation
CNN model
Augment metadata
Concat feature
Dense linear layer with BN/ReLU/dropout
Classification output
Patient metadata
Dense linear layer with BN/ReLU/dropout