Skin Cancer detection using ML

Names: Tarandeep Singh, Prayag Parashar, Praddume Attri,Udbhav Yadav

Group number: 16

ML-mid sem project

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY DELHI

Motivation

Detecting cancer early is crucial for better patient outcomes, but traditional methods often lack precision. Machine learning (ML) can help by analyzing complex data and making more accurate predictions. This project aims to use ML to improve cancer detection, enabling earlier diagnosis and better treatment options.

The idea came from seeing how machine learning is advancing in medical diagnostics. Its potential to boost accuracy and efficiency through smart algorithms and data analysis has sparked interest in applying it to cancer screening.

Literature review

The paper "*Skin Cancer Detection Using Deep Learning – A Review*" highlights the potential of deep learning (DL), especially **convolutional neural networks (CNNs)**, in improving skin cancer detection. It discusses the limitations of traditional diagnostic methods, like biopsies, and the benefits of non-invasive DL approaches. The review focuses on melanoma, the most aggressive form of skin cancer, and details how CNNs analyze dermoscopic images for accurate diagnosis.

The paper also covers key datasets like ISIC and HAM10000, challenges like data imbalance, and DL models such as **ResNet**, **DenseNet**, and **Inception**. Techniques like transfer learning and evaluation metrics like accuracy and sensitivity are also explored.

However, challenges include model interpretability, data bias (favoring lighter skin tones), and clinical integration. Future directions include addressing these issues and exploring DL's role in telemedicine to improve early detection, especially in underserved regions. Overall, DL shows promise but needs further development for widespread use in healthcare.

Literature review

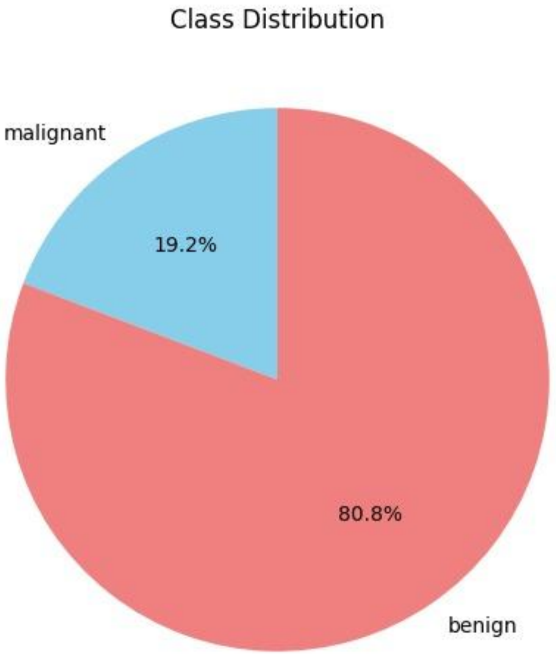
The paper "*A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis*" reviews AI's role in cancer research, focusing on machine learning (ML) and deep learning (DL) techniques. It analyzes 185 studies from 2009 to 2021, emphasizing how *convolutional neural networks* (**CNNs**) have improved cancer detection, especially in imaging like CT scans, MRIs, and X-rays, with applications in breast and lung cancer. Despite achieving a high accuracy of 98.4%, the paper highlights challenges such as data quality, model interpretability, and the need for broader research across populations to ensure AI's successful clinical integration.

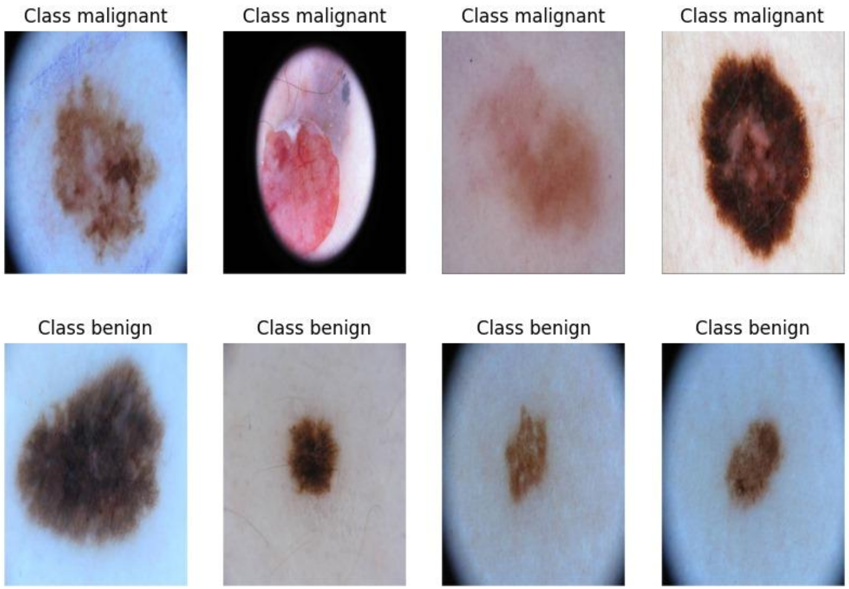
Dataset description

Link to the Dataset

The ISIC 2016 Original Dataset consists of images used for skin lesion analysis, specifically aimed at aiding the diagnosis of skin cancer. It includes various types of lesions categorized into different classes, such as melanoma, nevi, and seborrheic keratosis. The dataset provides high-resolution images, alongside annotations and labels for each lesion type, making it a valuable resource for training machine learning models in dermatology and medical image analysis.

Dataset description



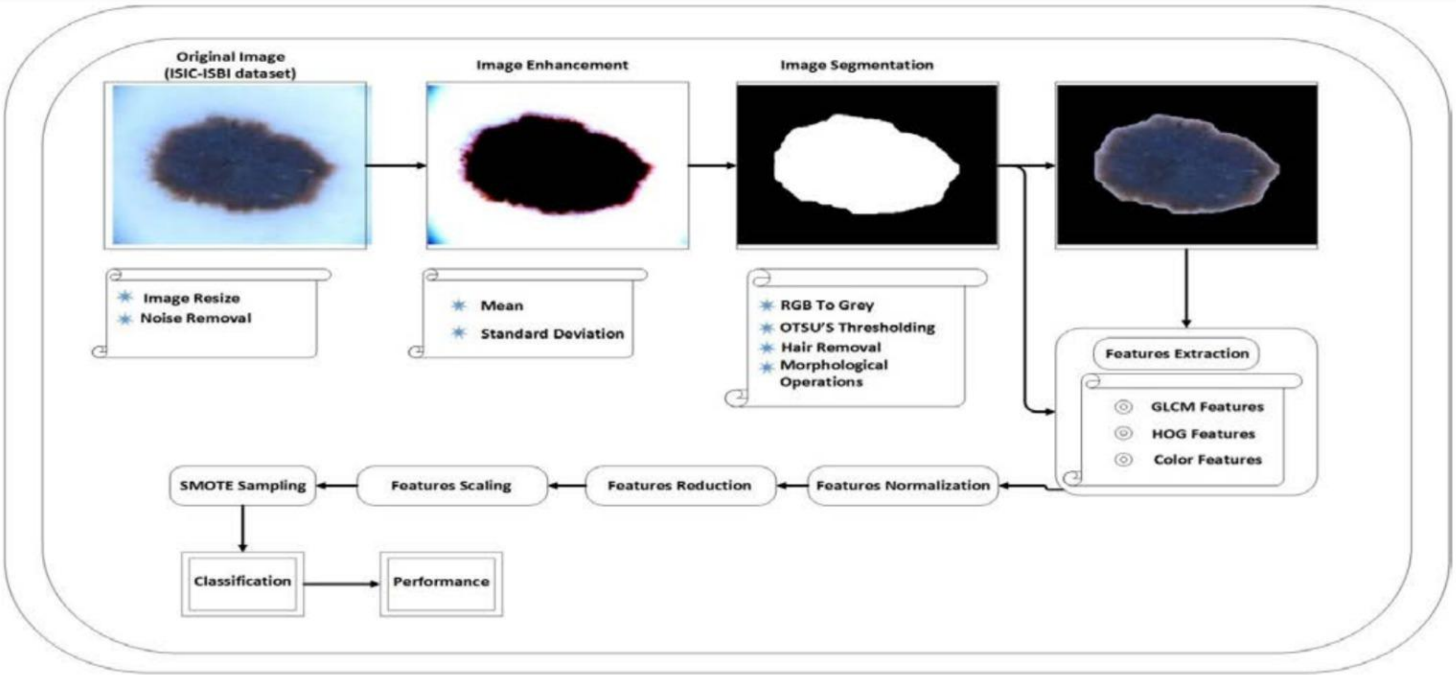


Methodology

In this paper, we propose a system that classifies dermoscopic images as benign or malignant by using image processing and machine learning. Steps include:

1. Pre-Processing
2. Contrast Stretching & Segmentation
3. Feature Extraction and Reduction
4. Data Balancing (SMOTE Sampling)
5. Feature Standardization, Scaling and Selection
6. Classification

Methodology

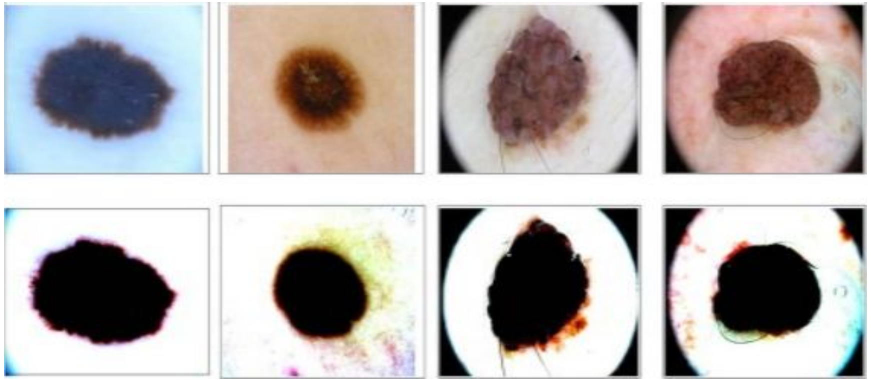


Pre-processing

- Applied a 3x3 median filter to remove 'salt and pepper' noise.
- Converted the image to grayscale for easier hair detection.
- Detected hair using bottom-hat filtering with a rectangular kernel.
- Restored image by inpainting the areas where hair was removed.

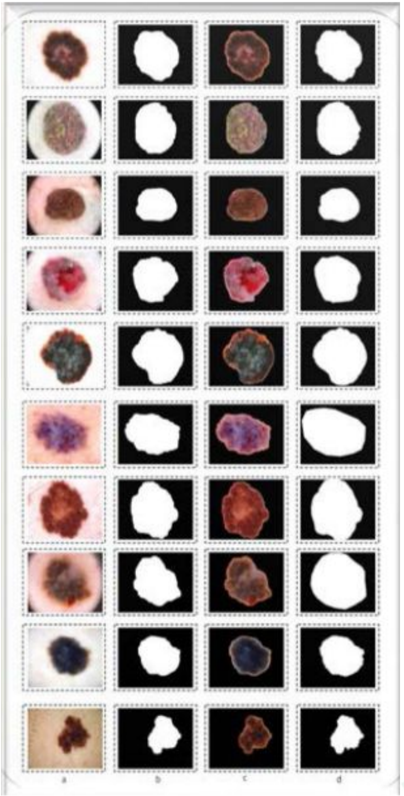
Contrast Stretching

- Splits the image into red, green, and blue channels.
- Enhances contrast by adjusting pixel values based on the mean and standard deviation for each channel.
- Merges the adjusted channels to create the final enhanced image.



Segmentation

- Converts the image to grayscale.
- Uses Otsu's method to automatically find the best threshold.
- Segments the image into foreground and background, creating a binary image.



Feature Extraction & Reduction

The process involves extracting features from segmented lesions, including 14 texture features via wavelet decomposition, 36 color features from RGB, HSV, and LAB color spaces, and shape features using Histogram of Gradients (HOG). Principal Component Analysis (PCA) reduces the HOG features from 3456 to 100. All features are combined into a sing

TABLE 1. EXTRACTED TEXTURE AND COLOR FEATURES

Texture features	Skewness	Mean	Contrast	Energy	Homogeneity	Standard
	Root Mean Square	Variance	Smoothness	Kurtosis	Correlation	Entropy
	Inverse Difference Movement					
Color Features	Mean R	Mean G	Mean Blue	Variance R	Variance G	Variance Blue
	Kurtosis R	Kurtosis G	Kurtosis	Skewness R	Skewness G	Skewness Blue
	Mean hue	Mean Saturation	Mean	Variance Hue	Variance	Variance Value
	Kurtosis H	Kurtosis S	Kurtosis V	Skewness H	Skewness S	Skewness V
	Mean L	Mean A	Mean B	Variance L	Variance A	Variance B

SMOTE Sampling

The dataset has a class imbalance problem, with 80% benign and 20% malignant samples. To balance the data, the SMOTE technique is used, creating an equal split (50% benign, 50% malignant) for both training and testing datasets

Results/Analysis/conclusion

SVM Classifier:

- Best Cross-validation accuracy: 96.14%
- Test accuracy: 80.56%
- Benign cases: Precision 0.82, Recall 0.97
- Malignant cases: Precision 0.50, Recall 0.11

Random Forest Classifier:

- Overall accuracy: 75%
- Benign predictions: 81% correct
- Malignant predictions: 25% accurate

Logistic Regression Classifier:

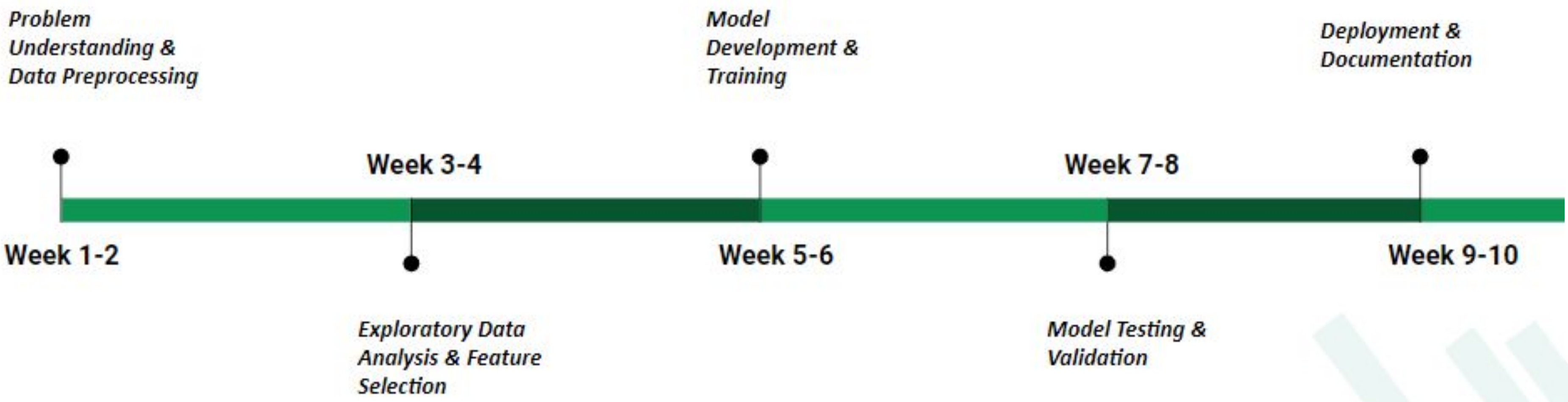
- Accuracy: 72.22%
- Benign cases: Precision 0.86, Recall 0.78
- Malignant cases: Precision 0.35, Recall 0.49

Classifier	Test Accuracy
SVM	80.56%
Random Forest	75%
Logistic Regression	72.22%

Conclusion

This study introduces a machine learning method for classifying skin cancer through image processing. It enhances dermoscopic images by adjusting contrast and applies OTSU thresholding for segmentation. Features like shape, color, and texture are extracted, with shape features simplified using PCA. The SMOTE technique addresses class imbalance in the ISIC dataset by generating synthetic samples. Finally, a new feature selection method using wrapper techniques is tested with a Random Forest classifier, yielding improved results on the ISIC-ISBI 2016 dataset compared to other classifiers.

Timeline


Problem Understanding & Data Preprocessing
Model Development & Training
Deployment & Documentation
Week 3-4
Week 7-8
Week 1-2
Week 5-6
Week 9-10
Exploratory Data Analysis & Feature Selection
Model Testing & Validation

Individual team members' contributions

- Prayag Parashar (2022377): PPT creation, Programming
- Tarandeep Singh (2022536): Report Creation, Programming
- Praddume Attri (2022358): Report Creation, Programming
- Udbhav Yadav (2022549): PPT creation, Programming

Thank You