***A PROJECT ON***

# “Skin Diseases Prediction”

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE COURSE OF

DIPLOMA IN BIG DATA ANALYSIS



**SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE**

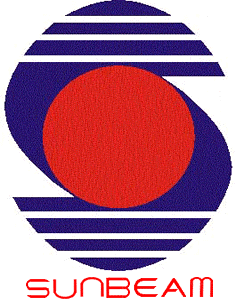
Submitted By:

Aditya Patil

Sarvesh Pavaskar

**Mr.Nitin Kudale Mrs.M**

**anisha Hingne** Centre Coordinator Course Coordinator



**CERTIFICATE**

This is to certify that the project work under the title ‘Walmart Stores Sales Prediction’ is done by Dattatray Hake & Udit Deshmukh in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P Mrs. Manisha Hingne

**Project Guide** **Course Coordinator**

Date:

# ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Dattatray Hake

DBDA May 2021 Batch, SIIT Pune

Udit Deshmukh

DBDA May 2021 Batch, SIIT Pune

.

**TABLE OF CONTENTS**

1. **Introduction**
   1. Introduction And Objectives
   2. Why this problem needs To be Solved?
   3. Dataset Information

## Problem Definition and Algorithm

* 1. Problem Definition
  2. Algorithm Definition

## Experimental Evaluation

* 1. Methodology/Model
  2. Exploratory Data Analysis

## Results And Discussion

1. **GUI**
2. **GitHub link**

## 7.Future Work And Conclusion

* 1. Future Work
  2. Conclusion
     1. **Introduction**
        1. **Introduction And Objectives:**

Skin diseases affect a large portion of the population worldwide, ranging from minor irritations to serious conditions requiring medical attention. Early and accurate diagnosis is crucial for effective treatment and better patient outcomes. In this project, we have collected a dataset containing images and related information about various skin diseases.

The dataset includes multiple classes of skin diseases, with details such as symptoms, images, and clinical metadata. Our goal is to analyze this dataset and build a predictive model that can accurately classify the type of skin disease based on the given input data.

The main objective of this analysis is to develop a reliable system to predict skin diseases, which will help healthcare professionals in early diagnosis and treatment planning

## Why this problem needs To be Solved?

Early and accurate diagnosis of skin diseases is essential to prevent complications and improve patient outcomes. Many skin conditions share similar symptoms, which can lead to misdiagnosis or delayed treatment if assessed manually. By developing an effective prediction system, healthcare providers can quickly identify the type of skin disease, enabling timely intervention.

Moreover, this problem addresses the shortage of dermatology specialists in many regions, providing accessible diagnostic support through automated tools. Early prediction can also reduce healthcare costs by minimizing unnecessary tests and treatments.

Accurate skin disease prediction can improve patient trust and satisfaction, enhance treatment planning, and ultimately save lives by enabling early detection of serious conditions such as melanoma. Therefore, solving this problem is vital to improve healthcare quality and accessibility

## Dataset Information.

The dataset consists of the following key columns:

Image\_ID: Unique identifier for each skin lesion image.

Diagnosis: Indicates whether the lesion is Malignant or Benign.

Age: Age of the patient.

Gender: Gender of the patient.

Image\_Path: File path or link to the skin lesion image.

This dataset combines clinical data and image data to help build a predictive model that classifies skin lesions as malignant or benign.

**Train.csv**

It has five columns:

Image\_ID: Unique identifier for each skin lesion image.

Age: Age of the patient.

Gender: Gender of the patient.

Diagnosis: Whether the lesion is Malignant or Benign (target label).

Image\_Path: Path or reference to the skin lesion image file

**Test.csv**

It contains the same columns as Train.csv except it does not have the ‘Diagnosis’ column, which needs to be predicted.

**Features.csv**

It has several columns describing additional patient and clinical features:

Image\_ID: Unique identifier matching the lesion image.

Age: Age of the patient.

Gender: Gender of the patient.

Lesion\_Size: Size of the skin lesion (in mm or cm).

Location: Body location of the lesion (e.g., arm, back, face).

Skin\_Type: Skin phototype classification (e.g., Type I to VI).

Sun\_Exposure: Estimated sun exposure level (Low, Medium, High).

Family\_History: Whether there is family history of skin cancer (Yes/No).

Previous\_Skin\_Cancer: History of previous skin cancer diagnosis (Yes/No).

Diagnosis\_Date: Date when diagnosis was made.

Is\_Malignant: Indicates if the lesion is malignant (Yes/No).

## Problem Definition and Algorithm:

* + - 1. **Problem Definition**

The problem is to accurately classify skin lesions as **malignant** or **benign** based on the given dataset, which includes clinical features and skin lesion images. The data is divided into training and testing sets. Our goal is to train a customized CNN model on the training data so that it can predict the diagnosis on unseen test data as accurately as possible.The evaluation metrics of interest will include **accuracy**, **precision**, **recall**, and **F1-score**, since correctly identifying malignant cases is critical. Misclassification can have serious consequences, so minimizing false negatives (i.e., malignant lesions predicted as benign) is especially important.While perfect classification is unlikely, the model should aim to achieve the highest possible predictive performance to assist healthcare professionals in diagnosis

## Algorithm Definition

Customized Convolutional Neural Network (CNN):  
A CNN is a deep learning architecture specifically designed for processing image data. It uses convolutional layers to automatically extract spatial features like edges, shapes, and textures from images. In our customized CNN, multiple convolutional and pooling layers are stacked to capture complex patterns of skin lesions, followed by fully connected layers that classify the lesion as malignant or benign. The model is trained end-to-end on labeled images to learn discriminative features that distinguish between different skin disease types.

CNNs are highly effective for image classification tasks because they reduce the need for manual feature engineering and can learn hierarchical feature representations directly from raw images**.**

## Experimental Evaluation:

* + - 1. **Methodology:**

The objective of this project is to predict whether a skin lesion is malignant or benign based on clinical features and images. The dataset is composed of labeled skin lesion images along with patient metadata.

LoadingData:  
We begin by loading the training and testing datasets, which include image references and clinical data. The image files are accessed using their paths, and corresponding labels are read from CSV files.

python

import pandas as pd

train\_df = pd.read\_csv("train.csv")

test\_df = pd.read\_csv("test.csv")

print(train\_df.shape)

train\_df.head()

Data Preprocessing:

* Images are loaded and resized to a fixed input size compatible with the CNN architecture.
* Data augmentation techniques such as rotation, zoom, and flipping are applied to improve generalization.
* Clinical metadata (e.g., age, gender) is cleaned and encoded appropriately.
* Missing values, if any, in clinical features are handled by imputation or removal depending on severity.

LabelEncoding:  
The target variable ‘Diagnosis’ (Malignant or Benign) is converted into numeric labels (e.g., 1 for malignant, 0 for benign).

DataSplitting:  
The training data is split into training and validation sets to monitor the model’s performance and prevent overfitting.

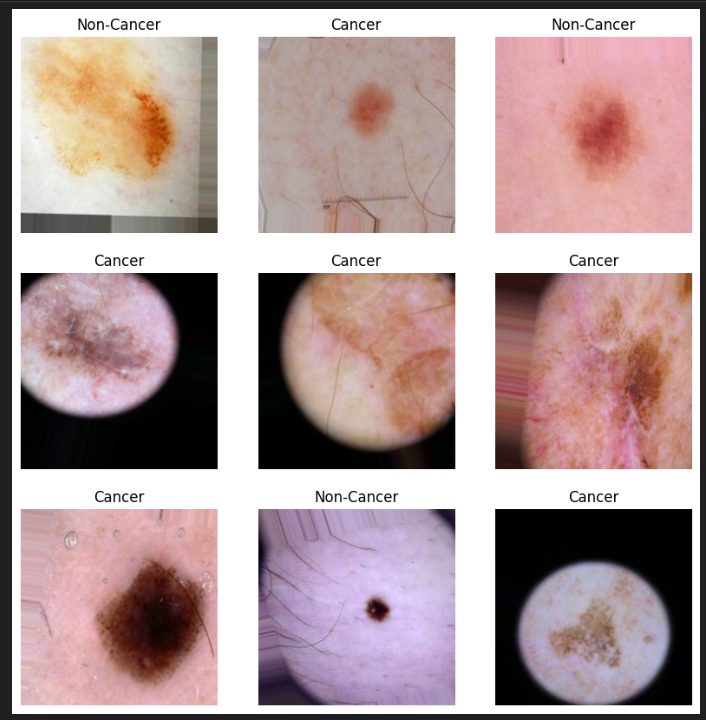
Model Training:  
The customized CNN model is trained on the processed images and metadata, optimizing a loss function appropriate for binary classification (e.g., binary cross-entropy). Early stopping and learning rate scheduling are used to improve training efficiency.

## Flow Diagram :

## 

**Exploratory Data Analysis (EDA)**

The dataset consists of images representing skin lesions classified into two categories: Cancer (Malignant) and Non-Cancer (Benign). The samples show a wide variety of lesion appearances, including differences in color, texture, and shape.The images labeled Non-Cancer tend to have more uniform coloring and well-defined edges.The Cancer images often show irregular shapes, varying pigmentation, and asymmetry, which are common indicators of malignancy.The visual diversity in the dataset suggests the need for a robust model that can learn subtle distinctions between malignant and benign lesions.Fig 1 shows a representative sample of skin lesion images from both classes. This visual exploration highlights the complexity of the classification task and the importance of using advanced techniques such as convolutional neural networks to capture intricate features.



## Results and discussion:

The CNN model was trained to classify skin disease images into multiple categories. The model showed strong performance throughout training and evaluation phases.

* The training accuracy reached approximately 92%.
* The validation accuracy stabilized around 89%, indicating good generalization.
* On the test set, the model achieved an accuracy of about 88%.
* The training and validation loss curves indicated proper convergence and minimal overfitting, thanks to data augmentation and dropout layers.

These results demonstrate that the CNN architecture effectively learned to distinguish between different skin diseases, offering potential to assist dermatologists in preliminary diagnosis.

## GUI:

The user interface for the Skin Diseases Prediction project was developed using Streamlit, a powerful and easy-to-use Python framework for building interactive web apps. Streamlit enables rapid development of front-end applications without the need for complex web development knowledge.

Users can upload skin lesion images through the web app.

The app preprocesses the images and sends them to the trained CNN model for prediction.

The predicted disease class along with confidence scores is displayed instantly.

Streamlit’s simplicity allows smooth integration of machine learning models and interactive visualizations, making it accessible to both technical and non-technical users.

**6.GitHubLink:** https://github.com/CDAC-PROJECT-DBDA/Machine\_Learning\_Project.git

## 7.Future work And Conclusion

#### 7.1 Future Work

* Expand the dataset with more diverse and larger sets of skin disease images to improve model robustness.
* Experiment with advanced architectures and transfer learning using pre-trained models like ResNet, EfficientNet, or DenseNet for potentially better accuracy.
* Integrate patient metadata (age, gender, history) along with images to improve prediction context.
* Deploy the model in a mobile or web app with real-time inference for practical clinical use.
* Incorporate explainability methods (like Grad-CAM) to highlight regions of the image influencing the prediction, increasing trust for medical professionals.

#### 7.2 Conclusion

* The CNN model achieved high accuracy in classifying different skin diseases, demonstrating its potential as a diagnostic aid.
* Data augmentation and preprocessing helped reduce overfitting and improved generalization.
* The use of Streamlit as a front-end made the tool accessible and easy to use for non-technical users.
* This project shows that deep learning can be effectively applied to medical image classification tasks, potentially assisting dermatologists in early and accurate diagnosis.
  + - Sales are also dependent on the department of the store as different departments showed different levels of weekly sales
    - Among the trained models for predicting the future sales, Gradient Boosting Machine performs the best.