

G H Patel College of Engineering & Technology

**(The Charutar Vidya Mandal (CVM) University)**

**New V. V. Nagar**

# DEPARTMENT OF COMPUTER ENGINEERING

**AI/ML Report**

**on**

***Cancer Detection System***

# Submitted By

**Name of Student :** Brij Shah

**Enrolment Number :** 12302130503007

**Name of Student :** Dev Somani

**Enrolment Number :** 12302130503004

# Guided By

**Prof. Priyang Bhatt**

# Artificial Intelligence & Machine Learning (202046702)

**A.Y. 2024-25 EVEN TERM**



# CERTIFICATE

This is to certify that the Mini Project Report titled **“Cancer Prediction System Using Machine Learning”** has been carried out by **Brij Shah (12302130503007)** under guidance in partial fulfilment for the Degree of Bachelor of Technology in **Computer Science & Design, 6th Semester**, G H Patel College of Engineering & Technology, CVM University, New Vallabh Vidyanagar during the academic year 2024-25.

Prof. Priyang Bhatt Dr. Sudhir Vegad

Internal Guide Head of Department



# CERTIFICATE

This is to certify that the Mini Project Report titled **“Cancer Prediction System Using Machine Learning”** has been carried out by **Dev Somani (12302130503004)** under guidance in partial fulfilment for the Degree of Bachelor of Technology in **Computer Science & Design, 6th Semester**, G H Patel College of Engineering & Technology, CVM University, New Vallabh Vidyanagar during the academic year 2024-25.

Prof. Priyang Bhatt Dr. Sudhir Vegad

Internal Guide Head of Department

**Title of the Project**

**Cancer Prediction System Using Machine Learning**

**Objective**

The objective of this project is to predict the likelihood of a patient having cancer based on clinical attributes. The system uses a supervised machine learning model trained on health-related tabular data to classify whether a patient has cancer or not. The end goal is to assist in early diagnosis and treatment planning using interpretable models and metrics.

**Dataset Used**

* **Source**: Breast Cancer Wisconsin (Diagnostic) Data Set or other relevant datasets (e.g., lung cancer or colon cancer datasets)
* **Files**:
  1. cancer\_images/: Directory containing medical image data, typically in formats such as JPG, PNG, or DICOM, for classification..
* **Key Features:**
  + Tumor Size: Size of the tumor detected in the image
  + Shape: Shape of the tumor (regular, irregular)
  + Density: Density of the tissue
  + Age of Patient: The patient's age
  + Diagnosis: 0 (Benign) or 1 (Malignant)

**Model Chosen**

1. Convolutional Neural Network (CNN) or Pretrained Models (e.g., ResNet34)
2. **Data Preprocessing**:
   * Images resized and normalized.
   * Augmentation techniques like rotation, scaling, flipping to improve model generalization.
   * Split data into training, validation, and test sets.
3. **Model Training:**
   * Trained a deep learning model (e.g., CNN or ResNet34) for binary classification (Benign or Malignant).
4. **Model Saving:**
   * Trained model saved as .pth (PyTorch model) for later use and deployment.

**Performance Metrics**

* **Classification Report:**
  + Shows precision, recall, and F1-score for each class (Benign and Malignant)..
* **Confusion Matrix:**
  + Provides insights into false positives and false negatives.
* **Example Metrics:**
  + Accuracy: ~77%
  + Precision (Diabetic): ~76%
  + Recall (Diabetic): ~72%

**Challenges & Learnings**

**Challenges:**

1. **Data Imbalance:**
   * Many cancer datasets (especially benign vs malignant) suffer from an imbalance, where benign cases far outnumber malignant ones.
2. **Image Quality Variability:**
   * Medical images can vary greatly in quality and resolution, which posed challenges in preprocessing and standardizing the data for training. .
3. **Model Interpretability:**
   * Understanding which features contributed most to predictions.

**Learnings:**

1. **Data Augmentation::**
   * Augmenting medical image data was essential to improve model robustness and accuracy..
2. **Transfer Learning:**
   * Transfer learning using pretrained models like ResNet34 improved accuracy significantly when dealing with medical images.
3. **Visualization:**
   * Interpreting feature importance, class distribution, and confusion matrices.

**Tools & Libraries**

* **Python:** Primary programming language
* **Libraries:** PyTorch, TorchVision, Pandas, NumPy, Matplotlib, Scikit-learn, OpenCV, Joblib
* **IDE:** Jupyter Notebook for EDA and prototyping
* **Environment:** Virtualenv or conda environment for isolation