BACHELOR OF COMPUTER ENGINEERING

(AFFILIATED TO JNTUA - Jawaharlal Nehru Technological University Anantapur)
MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE (MITS)
Angallu, Madanapalle – 517325, Chittoor District, Andhra Pradesh, India

PROJECT REPORT

ON

CANCER AND DEATH DETECTION

Submitted by

PINJARI AYESHA

USN : 24695A0501

Under the guidance of

- Mr. AKASH V

Project report submitted in partial fulfilment of the requirements of the Fourth Semester B.Tech
Madanapalle Institute of Technology & Science (MITS)

JULY-2025

TABLE OF CONTEXT

S.NO	CONTENT
1	Introduction
2	Proposed System
3	Advantages and disadvantages
4	Software & Hardware Requirements
5	Conclusion

1. Introduction

The early detection of cancer and prediction of patient survival are among the most critical applications of data science in healthcare. This project presents a machine learning-based approach for classifying cancer presence and predicting patient mortality based on historical and clinical data. Using the Logistic Regression algorithm, the system learns from existing data to make reliable predictions about unseen cases.

A key innovation in this project is the introduction of a 'Risk Score', a custom-calculated value that indicates the likelihood of death or severe illness in cancer patients. This score provides a quantifiable measure of patient risk, supporting doctors in treatment prioritization and early intervention.

By analyzing features like tumor size, age, lymph node status, and survival history, the system supports faster, data-driven decisions. The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score, demonstrating its practical potential in improving diagnosis and patient care.

The objective of this work is to demonstrate the effectiveness of machine learning in healthcare analytics and to build a predictive system that can aid in early diagnosis, improve patient outcomes, and potentially reduce the burden on healthcare systems. The model's performance is evaluated using accuracy, precision, recall, and F1-score metrics to ensure its reliability for practical use.

2. Proposed System

System Overview

The project implements two primary predictive models:
- Cancer Detection: Predicts whether a patient is likely to have cancer.
- Death Prediction: Predicts the survival outcome (Living/Deceased) based on various clinical features.

System Components

1. **Data Collection**
   - Datasets Used:
     - Breast Cancer Wisconsin Dataset
     - METABRIC (Molecular Taxonomy of Breast Cancer)

   2. Data Preprocessing
   - Handling missing values
   - Label encoding for categorical features
   - Feature scaling using StandardScaler
   - Generation of the Risk Score as an additional derived feature

   3. Exploratory Data Analysis (EDA)
   - Visualizations (histograms, box plots, KDE plots)
   - Feature correlation analysis
   - Risk Score distribution vs. survival outcome

   4. Model Development
   - Algorithm Used: Logistic Regression
   - Dataset Split: train_test_split() used to create training and testing sets
   - Model Evaluation:
     - Accuracy Score
     - Precision, Recall, F1-Score
     - Confusion Matrix

   5. Special Feature: Risk Score

- A calculated metric representing the probability of death
- Visualized using seaborn to compare Risk Score across survival statuses
- Used to enhance interpretation of logistic regression predictions

6. Model Results
- Cancer Detection Accuracy: 97.36%
- Death Prediction Accuracy: 67.63%

2. Advantages and Disadvantages

Advantages

- Provides early, automated detection of cancer risk
- Risk Score helps triage and prioritize high-risk patients
- High accuracy for cancer detection ensures diagnostic reliability
- Logistic Regression offers interpretable results for clinical use
- Supports personalized medical decision-making

Disadvantages

- Death prediction is inherently more complex and less accurate
- Risk Score calculation is based on assumptions and may vary by case
- Accuracy depends heavily on data quality and feature selection
- Logistic Regression may underperform for non-linear relationships
- Requires strict data privacy and compliance due to medical sensitivity

4.Software & Hardware Requirements

Software Requirements

- Python: The programming language used for data analysis, machine learning,

and model development.

• Jupyter Notebook: An interactive development environment that allows for easy

prototyping, visualization, and documentation of data analysis workflows.

• NumPy: A fundamental library for numerical computing in Python, essential for

handling arrays and mathematical operations

• Pandas: A powerful library for data manipulation and analysis, particularly

useful for loading, cleaning, and preprocessing datasets.

• Scikit-learn: A machine learning library in Python that provides a wide range of

algorithms for classification, regression, clustering, and model evaluation.

Hardware Requirements

• Computer: A standard laptop or desktop computer with sufficient processing power and memory capacity to handle data analysis tasks.

• Processor: A multi-core processor (e.g., Intel Core i5 or higher) for faster computation of machine learning algorithms.

• Memory (RAM): At least 8 GB of RAM is recommended to handle large datasets and complex machine learning models efficiently.

.5. Conclusion

This project demonstrates the potential of machine learning in advancing healthcare by predicting cancer presence and patient mortality. Using the Logistic Regression algorithm, we developed predictive models that classify whether a patient is likely to have cancer and estimate survival outcomes based on clinical features. A key enhancement is the integration of a custom-calculated Risk Score, which quantifies the severity and likelihood of death, supporting better medical decision-making.

The analysis revealed that features such as tumor size, age at diagnosis, lymph node involvement, and survival history significantly impact both cancer detection and mortality prediction. The model achieved high accuracy in cancer classification (97%), while mortality prediction reached moderate accuracy (~68%), which is expected given the complexity of patient outcomes.

The Risk Score proved especially useful in visualizing and interpreting risk levels across patients, offering an intuitive way for healthcare providers to identify and prioritize high-risk individuals. This score adds value by making the model outputs more interpretable and actionable in clinical settings.

In practical terms, this system enables early diagnosis, supports prioritization of treatment, and has the potential to reduce mortality by facilitating timely intervention. While the current model focuses on logistic regression for interpretability, future work could explore ensemble methods or neural networks to enhance prediction accuracy further.

In conclusion, the successful use of machine learning and risk quantification in this project highlights the power of data-driven healthcare solutions. This project serves as a foundation for building more intelligent, accurate, and interpretable tools for early disease detection and patient care.