Cancer and Death Detection using Machine Learning

Pinjari Ayesha (24695A0501)

Guide: Mr. Akash V

MITS College, July 2025

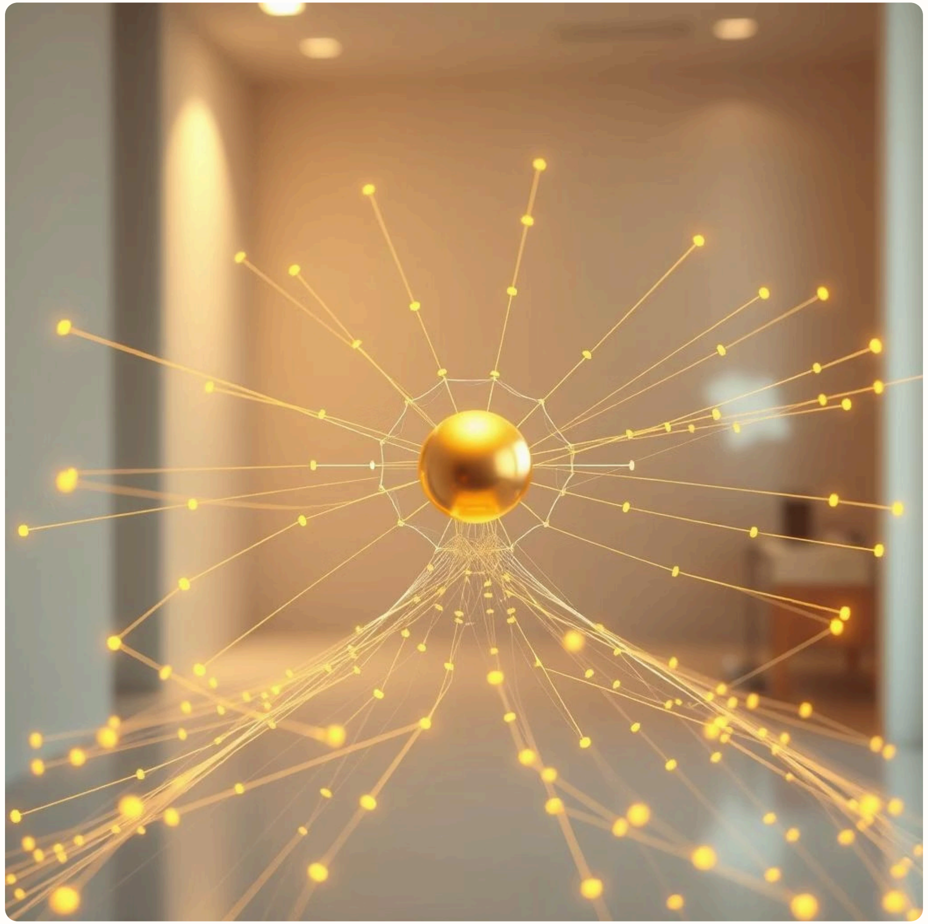
Introduction: Project Overview

Problem: Late diagnosis of cancer significantly reduces survival rates. Predicting the likelihood of death in patients is crucial for timely intervention.

Solution: Leveraging Machine Learning to enhance early detection and provide prognostic insights.

Project Objectives

- Develop robust cancer detection models.
- Build accurate death prediction models.
- Integrate these models for comprehensive patient insights.



Motivation: Why ML in Healthcare?

Enhanced Accuracy

ML algorithms process vast datasets, identifying subtle patterns often missed by traditional methods.

Early Intervention

Timely diagnosis improves treatment outcomes and patient survival rates significantly.

Clinical Support

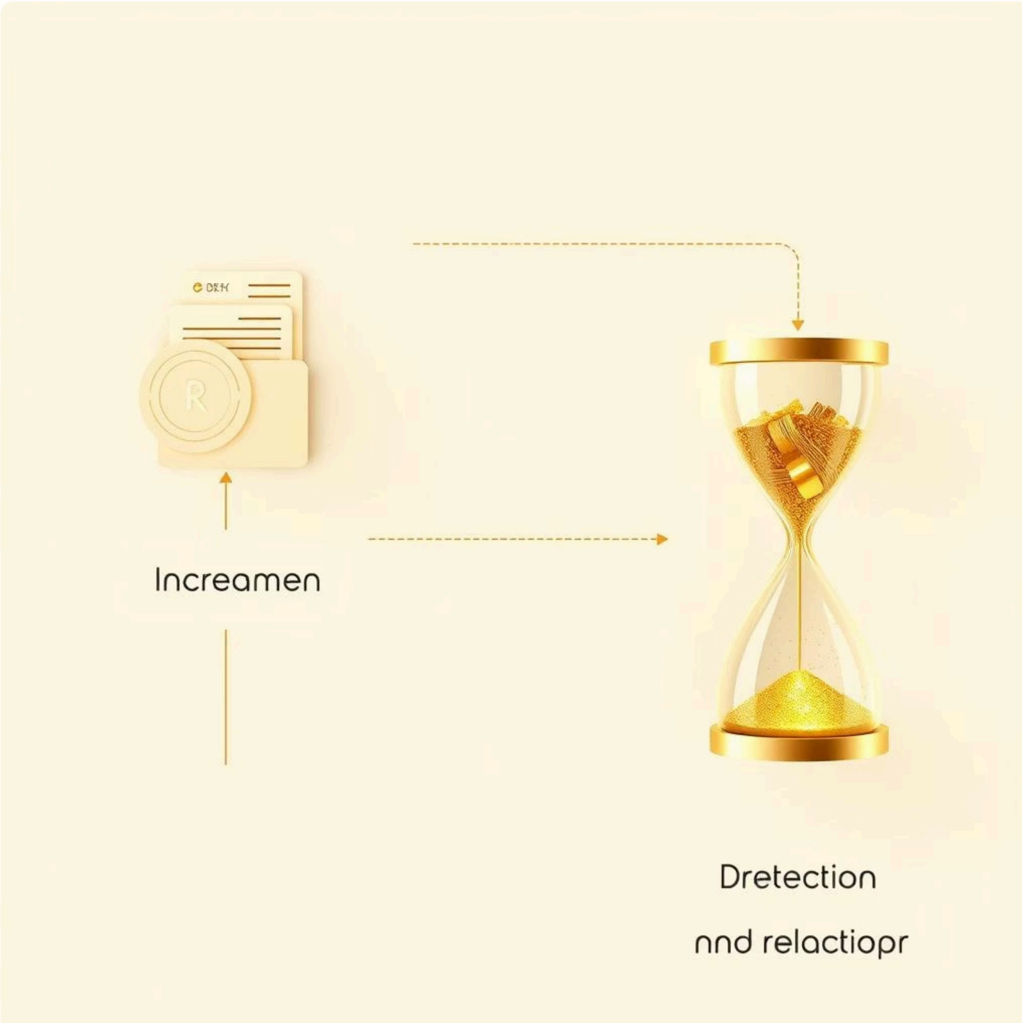
Provides decision-making support for clinicians, leading to more informed patient care plans.

Prognostic Insights

Predicts disease progression and patient outcomes, enabling proactive medical strategies.

> Late diagnosis presents critical challenges, underscoring the urgent need for advanced analytical tools in healthcare.

System Overview



1

Cancer Detection Model

Identifies the presence of cancer based on diagnostic features. Focuses on classification accuracy.

2

Death Prediction Model

Assesses the probability of patient mortality using various clinical and demographic factors. Aids in prognosis.

Two distinct, yet integrated, models work in tandem to provide a holistic view of patient health status and future risks.

Workflow Diagram

1

Data Collection

Gathering diverse medical datasets.

2

Preprocessing

Cleaning and transforming raw data.

3

EDA

Discovering patterns and insights.

4

Model Training

Building predictive algorithms.

5

Evaluation

Assessing model performance.

6

Results

Interpreting findings for clinical application.

Data Preprocessing & Risk Score

Raw medical data often contains inconsistencies and missing values. Rigorous preprocessing ensures model accuracy and reliability.

- ## Missing Value Handling

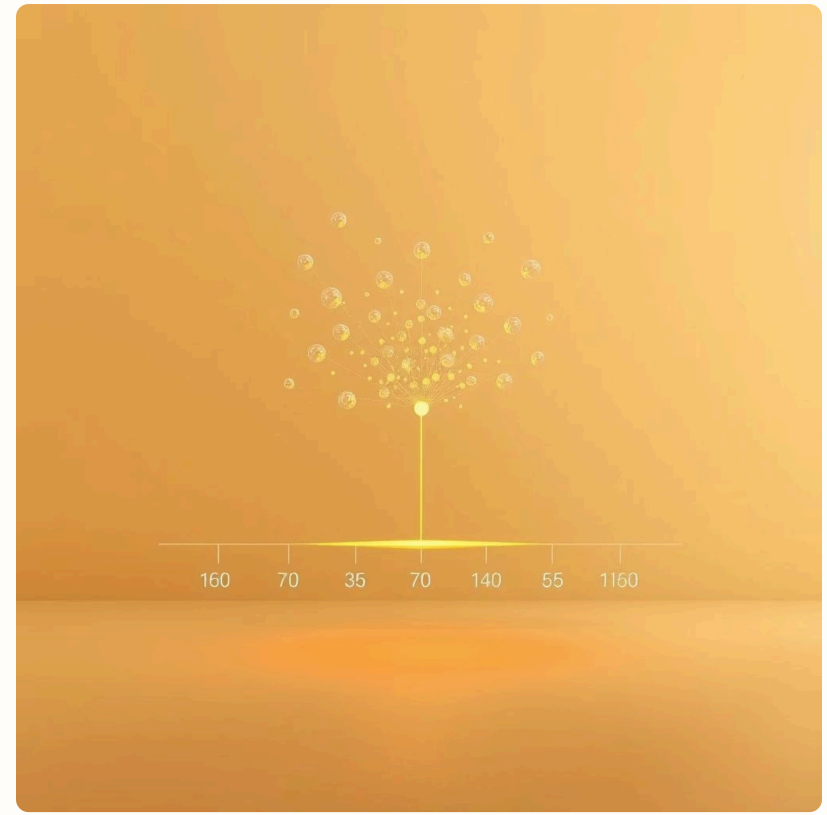
  Imputation techniques used to fill gaps in the dataset.

- ## Categorical Encoding

  Converting non-numeric data into a machine-readable format.

- ## Feature Scaling

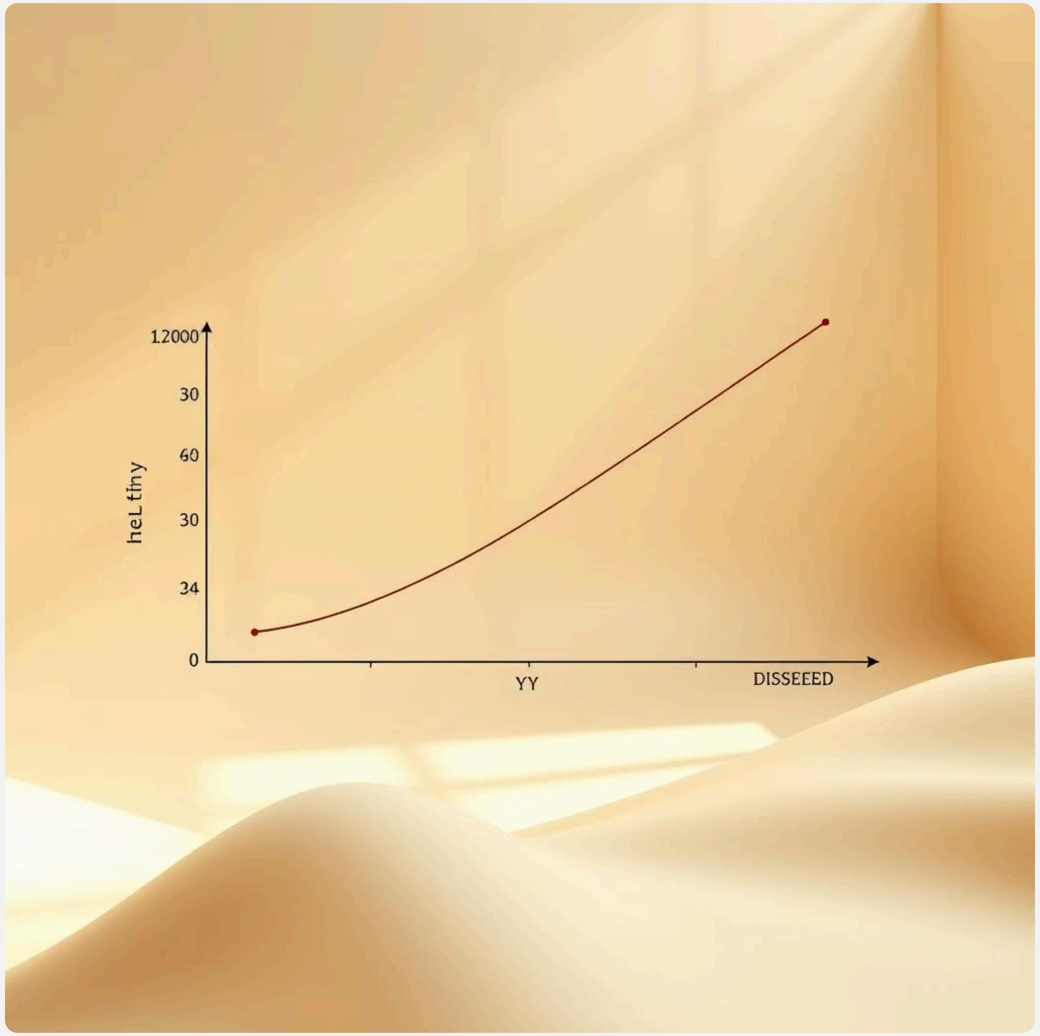
  Normalizing data ranges to prevent bias in model training.



Custom Risk Score Feature

A composite metric derived from multiple patient features, providing a unified indicator of health risk. This score aids in simplifying complex data for clinical interpretation.

Model Development: Logistic Regression



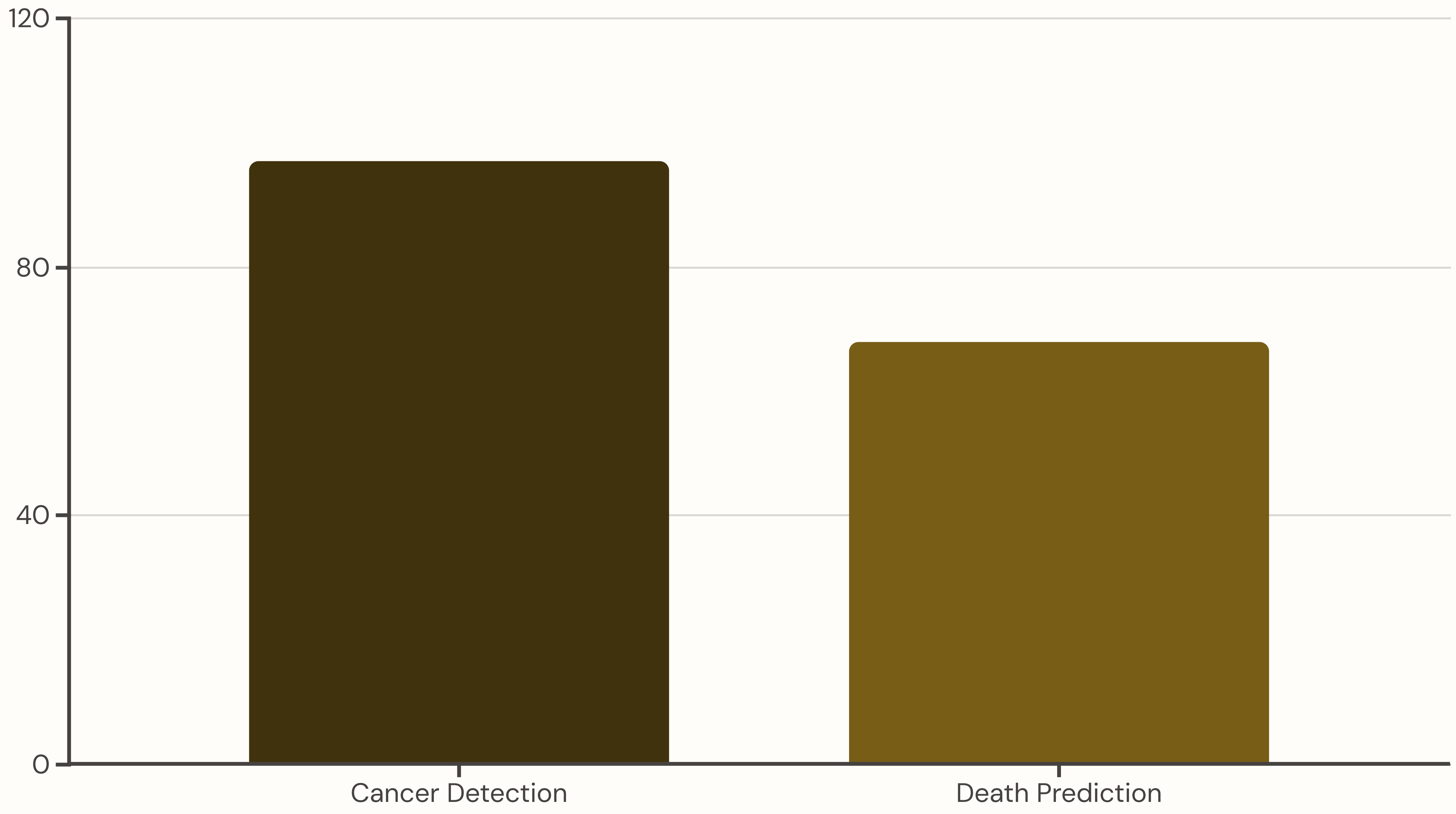
Why Logistic Regression?

- **Simplicity & Interpretability:** Easy to understand and explain, crucial for medical applications.
- **Binary Classification:** Ideal for predicting outcomes like cancer presence (yes/no) or death (yes/no).
- **Probabilistic Output:** Provides probability scores, offering more nuanced insights than just a class label.

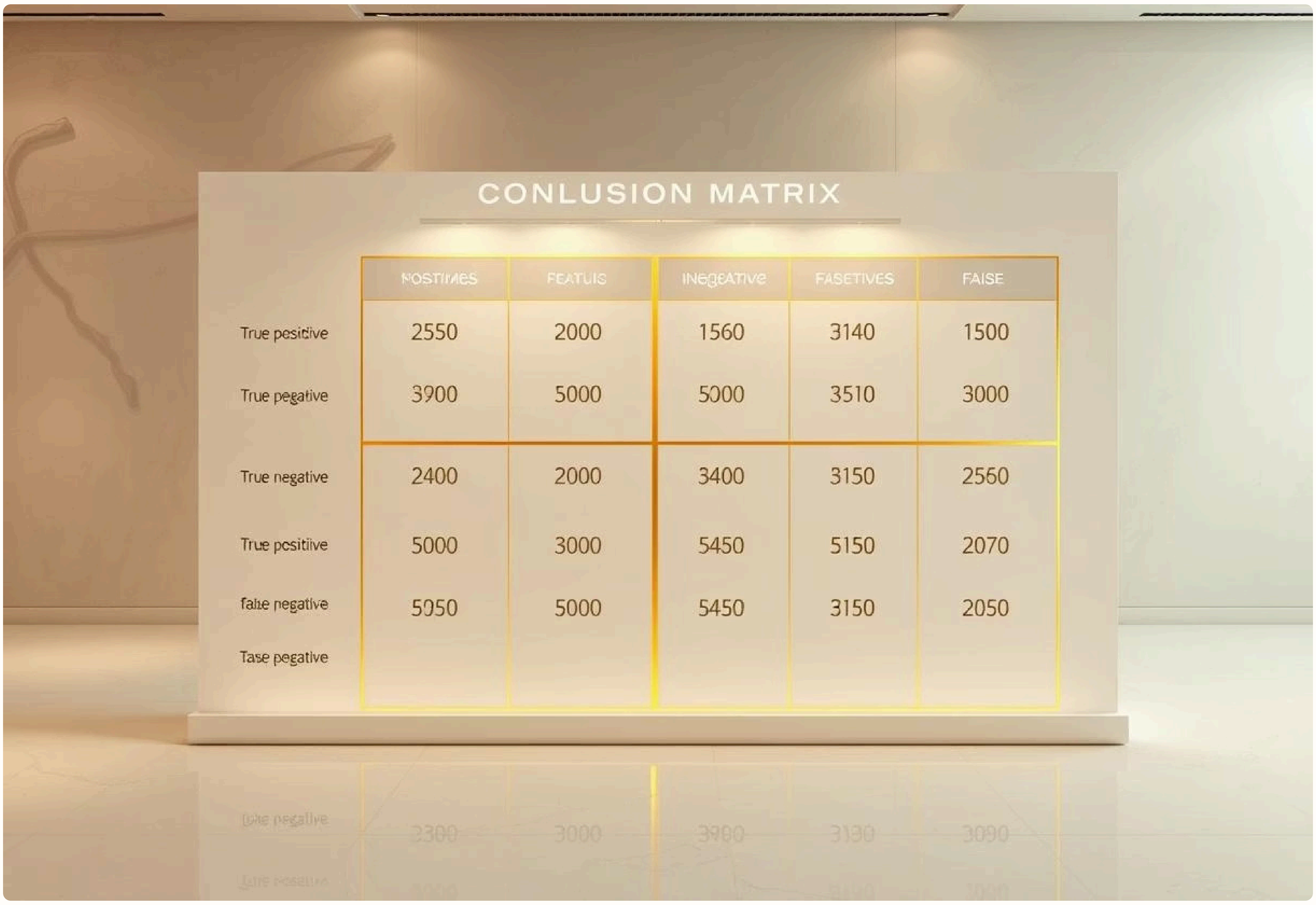
Train/Test Split

Data divided into training (70%) and testing (30%) sets to ensure robust model validation and prevent overfitting.

Results: Performance Metrics



Confusion Matrix Highlights:



- High True Positive Rate for cancer detection.
- Identified areas for improvement in death prediction, particularly in reducing false negatives.

Advantages vs. Disadvantages

Advantages

- Early detection capabilities.
- Improved diagnostic accuracy.
- Personalized risk assessment.
- Scalable for large datasets.
- Automated and consistent analysis.

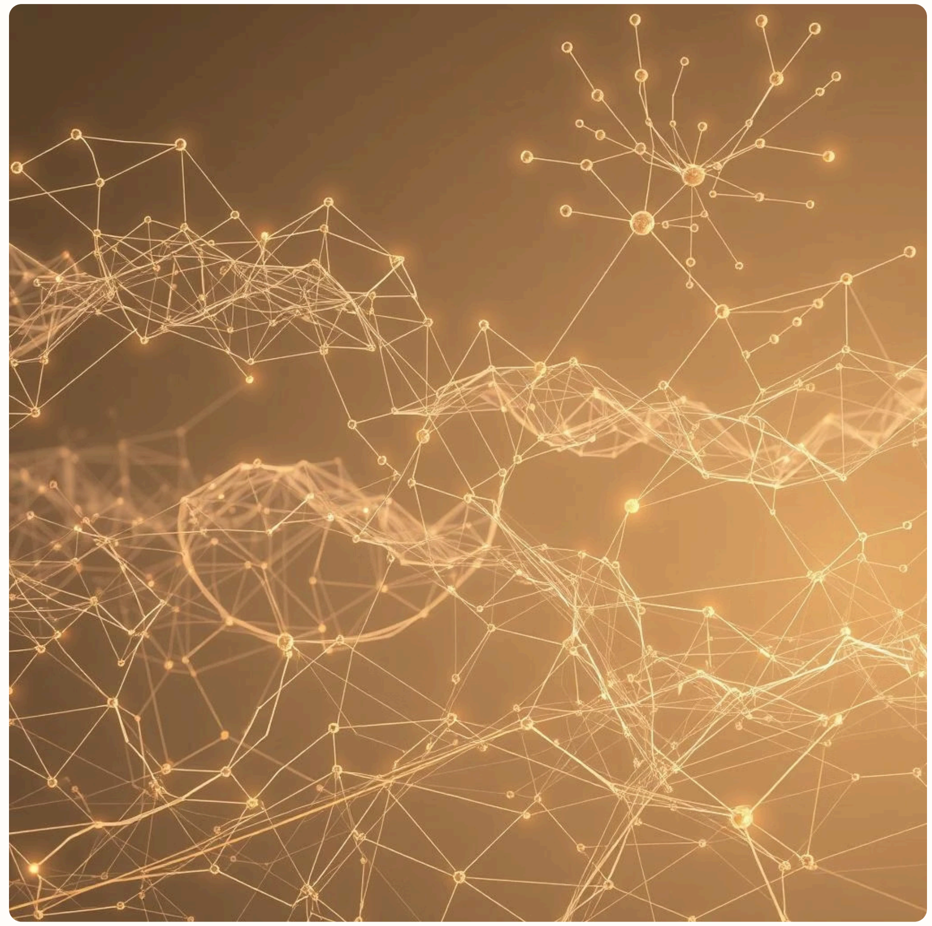
Disadvantages

- Dependency on data quality.
- Interpretability challenges (for complex models).
- Potential for algorithmic bias.
- Requires domain expertise for feature engineering.
- Generalizability across diverse populations.

Conclusion & Future Work

Risk Score Benefits

The custom **Risk Score** streamlines complex patient data into an intuitive metric, enhancing clinical decision-making and patient management.



Deep Learning

Explore neural networks for higher accuracy and automated feature extraction.

Ensemble Methods

Combine multiple models to improve overall predictive performance and robustness.

Clinical Validation

Future work includes rigorous testing with real-world clinical data for practical implementation.