CUFE-CHS 2024
Cairo University - Faculty of Engineering - Credit Hour System
QS World University Rankings RANKED 149 2024

Axillary Lymph Node Metastasis Prediction Using AI

Submitted By:
- AHMED OSAMA -1190415
- OMAR ESSAM - 1190332
- HASSAN SAMY – 1190025
- MAHREAL AYMAN - 1190491
- MONA MOHSEN-1190540

Under Supervision of :
Prof.Dr.Ibrahim Youssef

Abstracts:

Detecting lymph node involvement is crucial for breast cancer prognosis, but the standard Sentinel Lymph Node Biopsy (SLNB) method has potential side effects. This project developed a machine learning-based approach to detect metastasis, reducing the need for invasive surgery.
Using a dataset of 950 anonymized medical records from Baheya Foundation, we rigorously pre-processed the data and tested various models, identifying CatBoost as the most effective.
Among the models tested, the CatBoost model stood out, demonstrating superior performance with an accuracy of 83%, an F1 Score of 82%, and an Area Under the Curve (AUC) of 80%.
We created a user-friendly website as a clinical decision support system, helping doctors identify metastasis pre-operatively with visualization tools and filters. This system provides a reliable, non-invasive alternative to traditional methods.

Introduction:

Breast cancer is the most common cancer globally and the second leading cause of cancer deaths among women. In the U.S., over 4 million people have had breast cancer, and in Egypt, it accounts for 35% of female cancers. Metastasis has a 31% five-year survival rate in the U.S. Sentinel Lymph Node Biopsy (SLNB) assesses axillary lymph nodes but can have side effects like a high false-negative rate and arm or shoulder complaints. A study at Kasr Alainy University Hospital reported a false-negative rate of up to 17%.

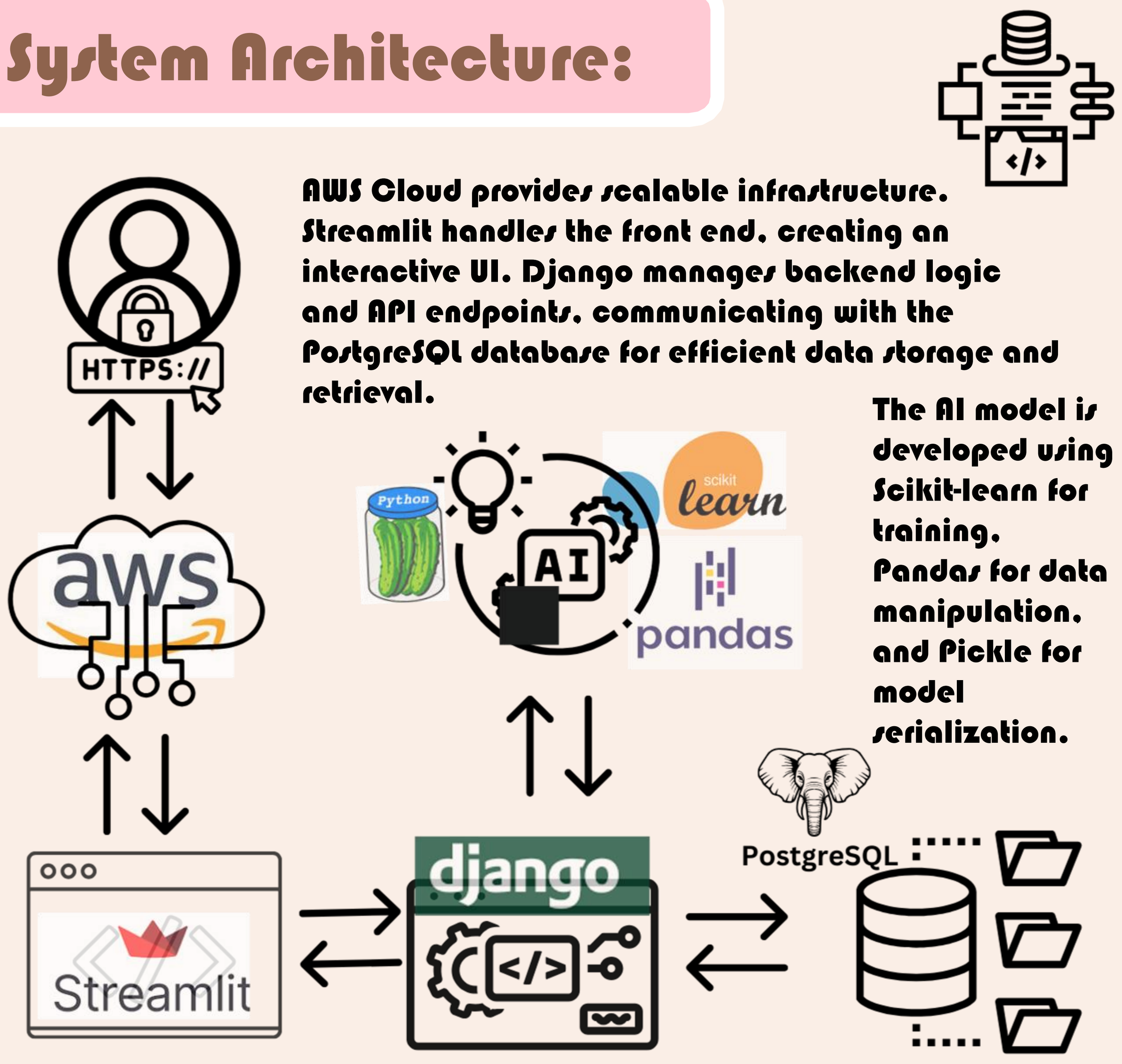
Objective:

- Create AI Web-Based Software to predict lymph node metastasis to Empower doctors and patients in the breast cancer journey.
  - Reduce unnecessary sentinel lymph node biopsies.
  - Improve preoperative diagnosis of axillary lymph node metastasis (ALNM).
  - Website Features:
    - Clinical decision support system.
    - Enhanced patient data management.
    - Advanced filtering for efficient patient care.

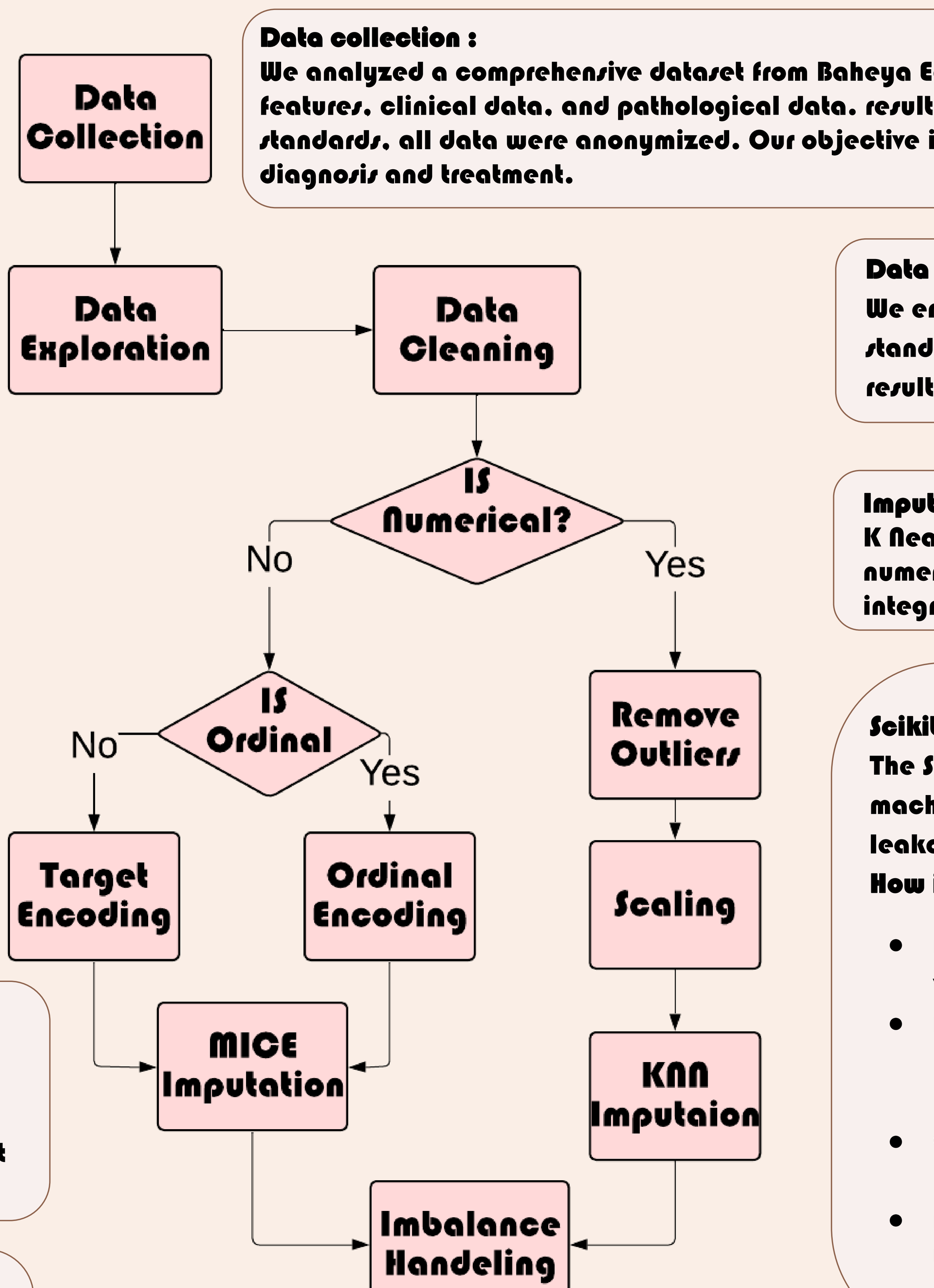
System Architecture:

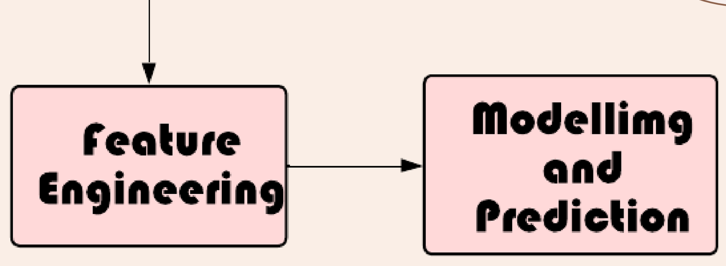
AWS Cloud provides scalable infrastructure. Streamlit handles the front end, creating an interactive UI. Django manages backend logic and API endpoints, communicating with the PostgreSQL database for efficient data storage and retrieval.

The AI model is developed using Scikit-learn for training. Pandas for data manipulation, and Pickle for model serialization.



Methodology:





Data collection :
We analyzed a comprehensive dataset from Baheya Egypt, comprising diverse data sources such as medical records, radiomic features, clinical data, and pathological data, resulting in a robust total of 950 samples from Baheya. To ensure strict privacy standards, all data were anonymized. Our objective is to leverage this multidimensional dataset to enhance breast cancer diagnosis and treatment.

$$F(x) = F_0(x) + \sum_{m=1}^{M}\sum_{i=1}^{n} f_m(x_i)$$

Exploration :
We performed an in-depth exploratory data analysis (EDA) to understand our dataset's characteristics, revealing a class imbalance (75% negative, 25% positive outcomes) and providing insights into data distribution and feature diversity

Data Cleaning :
We ensure data integrity by removing duplicates, standardizing formats, and correcting inconsistencies, resulting in a clean, uniform dataset ready for analysis.

F(x) is the overall prediction function.
Fo is the initial prediction function.
M is the number of trees in the model.
N is the number of samples in the training dataset.

Encoding :
Categorical features were transformed into numerical representations to ensure model compatibility and improve predictive performance. Ordinal encoding was applied to features with inherent order (e.g., T, N, M), while target encoding captured nuanced relationships between categorical features and the target outcome, enriching the dataset with valuable information for predictive modeling.

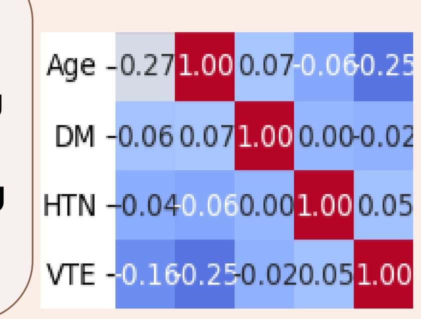
Imputation Techniques 2:
K Nearest Neighbors (KNN) Imputation: Estimated missing values in numerical features based on neighboring data points, preserving data integrity and accommodating non-linear patterns.

Scikit-learn Pipeline :
The Scikit-learn Pipeline integrates multiple preprocessing steps with a machine learning model, simplifying workflows and mitigating data leakage risks.
How it Works :

- **Preprocessing Steps:** Includes scaling, imputation, and categorical variable encoding, executed sequentially by transformer objects.
- **Model Fitting:** Transformed data is fed into the estimator object representing the machine learning model for pattern learning and prediction.
- **Chaining Steps:** All preprocessing and modeling steps are interconnected, executing sequentially within the pipeline.
- **Parameter Grid Search:** Combines with grid search for hyperparameter tuning, optimizing preprocessing and model parameters simultaneously.

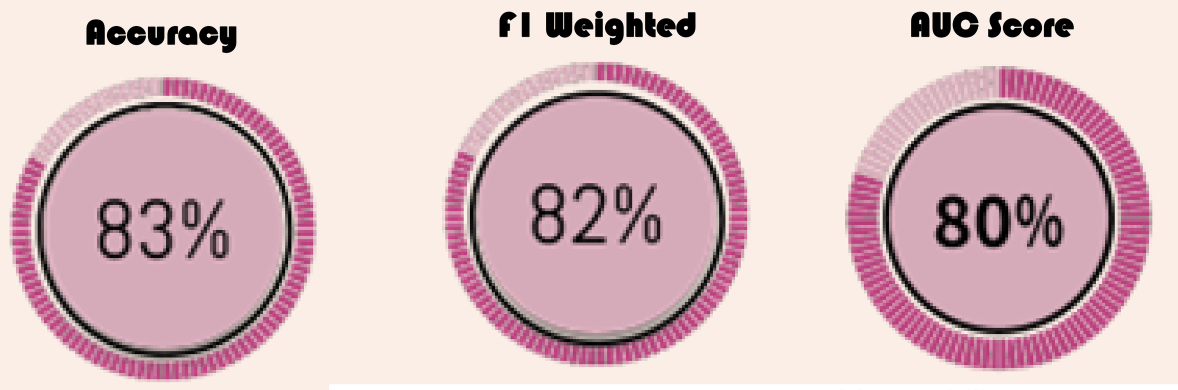
Imputation Techniques 1 :
Multiple Imputation by Chained Equations (MICE): Iteratively imputed missing values in categorical features by modeling each feature conditional on others, ensuring accurate and comprehensive dataset completion.

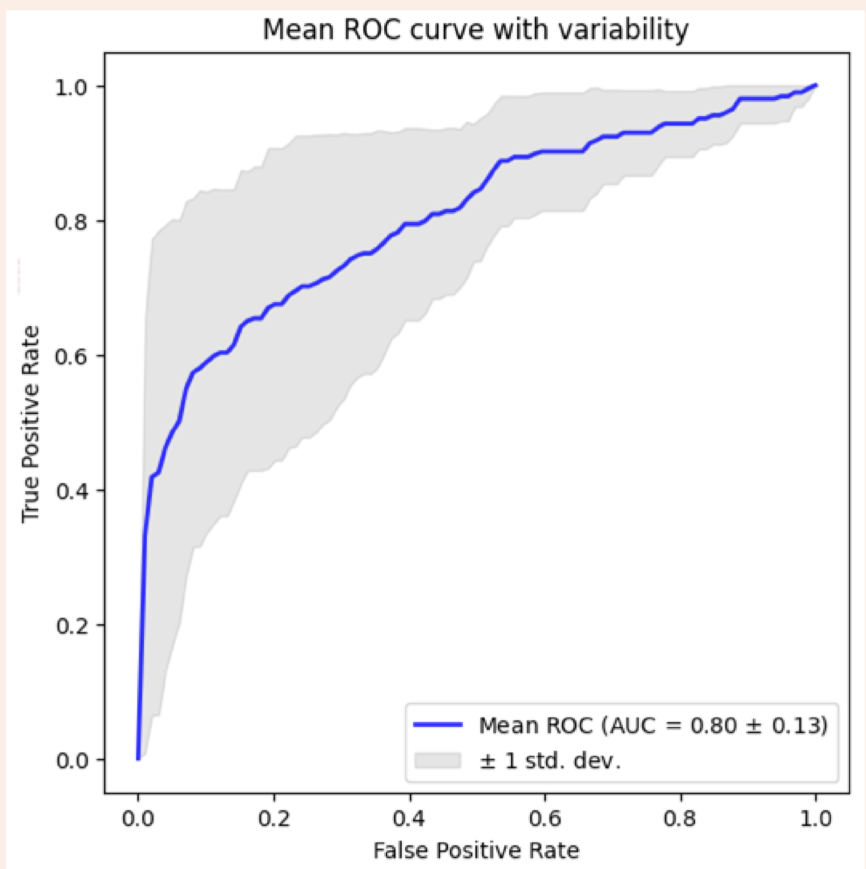
Feature Correlation Analysis :
Visualizing feature correlations using Seaborn's heatmap revealed relationships between features and the target variable, as well as inter-feature correlations. This analysis uncovered multicollinearity and dependencies, guiding us in eliminating redundant or less informative features.



CatBoost :
CatBoost efficiently handles categorical features, preserving patterns during conversion.

Results:





ALNM System: