Axillary Lymph Node Metastasis Prediction using AI

1st Ahmed Asaad
Healthcare Engineering and Management
Cairo University
Cairo, Egypt
ahmed.asaad01@eng-st.cu.edu.eg

2th Mahrael Salib
Healthcare Engineering
Cairo University
Cairo, Egypt
mahrael.salib02@eng-st.cu.edu.eg

3rd Mona El Batran
Healthcare Engineering
Cairo University
Cairo, Egypt
mona.hassan01@eng-st.cu.edu.eg

4th Omar El Ansary
Healthcare Engineering
Cairo University
Cairo, Egypt
Omar.ansary01@eng-st.cu.edu.eg

5th Hassan Oshaib
Healthcare Engineering
Cairo University
Cairo, Egypt
Hassan.abdelhameed02@eng-st.cu.edu.eg

6th Dr.Ibrahim Youssef
Healthcare Engineering
Cairo University
Cairo, Egypt
ibrahim.youssef@eng1.cu.edu.eg

***Abstract*—One of the most critical factors in breast cancer prognosis is lymph node involvement, significantly influencing treatment decisions. For detecting breast cancer metastasis , the standard method is the Sentinel Lymph Node Biopsy (SLNB) surgery, which involves assessing the status of axillary lymph nodes. However, SLNB has faced high criticism and can lead to potential side effects. Machine learning is a tool that can aid doctors in pre-operative diagnosis.This project develops a machine learning-based approach for detecting breast cancer metastasis. The model analyses a comprehensive set of pre-operative features, aiming to eliminate the need for invasive surgery and its associated side effects. This study was conducted in collaboration with Baheya Foundation, utilising a dataset of 950 anonymized medical records. The data underwent rigorous pre-processing steps to ensure quality. Subsequently, different machine learning models were employed to identify the most effective one. We developed a user-friendly website that acts as a clinical decision support system, enabling doctors to identify potential metastasis in breast cancer patients. The system also provides both visualisation tools and filters for patient results, empowering doctors to make more informed decisions. Our system delivers exceptional performance, exceeding the limitations of past research in this field. Notably, the CatBoost model achieved an accuracy of 83%, an F1 Score of 82%, and an AUC of 80%, demonstrating its effectiveness in prediction.The proposed system offers a reliable tool that can assist healthcare professionals.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Despite significant advancements in treatment and awareness, breast cancer remains the most common cancer globally [1]. Breast cancer constitutes the second leading cause of cancer mortality among women, accounting for approximately 15% of all female cancer deaths [2]. In the United States alone, there are currently over 4 million people with a history of breast cancer [3]. In Egypt, breast cancer accounts for approximately 35% of all female cancers according to data available as of 2022 [4]. Beyond the initial growth, the ability of breast cancer to spread, or metastasize, remains a formidable challenge in patient care. For patients diagnosed with distant breast cancer in United States, where the disease has metastasized to organs like the lungs, the five-year survival rate remains a concerning 31% [5]. This finding underscores the critical need for further research into metastasis, the process by which cancer cells spread from the primary tumor to other body parts.Sentinel lymph node (SLN) is the first node to receive the drainage directly from a tumor. The first studies on the role of the lymphatic system in the spread of cancer cells and metastasis were performed in the 1950s. SLNB was first reported in 1960 but took approximately 40 years to come into general practice following reports of good outcomes in patients with melanoma [6]. SLNB has become and remains the standard method for detecting breast cancer metastasis. However, SLNB comes with a range of potential side effects that can impact patients' recovery. It has faced criticism due to its high false negative rate which can sometimes be around 15% [7-8]. A study conducted at Kasr Alainy University Hospital found a false-negative rate for SLNB as high as 17% [9]. In addition, it's important to note that around 25% of women who undergo this procedure may still experience wound infection, delayed healing, or pain [10]. Among patients who reported arm and shoulder complaints following SLNB, 26.2% experienced loss of strength, 19.5% reported limitations in range of motion, and about 38.1% of the patients were treated by a physical therapist [11]. This study utilizes machine learning to predict whether a patient has metastasis, aiming to eliminate the side effects associated with the invasive surgery. It offers a tool to help doctors identify patients who may not require a sentinel lymph node biopsy, thereby minimizing the number of such procedures performed. Specifically, patients identified by model as having a low risk of metastasis may avoid the need for surgery. Our website acts as a clinical decision support system, that enhances patient data management by offering doctors a suite of functionalities. Doctors can create individual patient records and predict their metastasis outcome.The system facilitates secure storage and

retrieval of patient data through a dedicated database.The system empowers doctors with advanced filtering capabilities within their patient database. Breast cancer patients face a difficult journey. This motivated us to develop a tool that could empower both doctors and patients.

II. Literature Review

The table below presents a literature review of the studies that focus on prediction of breast cancer metastasis using AI utilizing clinicopathological features:

TABLE I
COMPARISON WITH EXISTING LITERATURE

Paper	Year	Method	Accuracy	AUC
[12]	2024	Deep Learning	75%	74%
[13]	2023	Machine Learning	-	76%
[14]	2022	Machine Learning	-	70%
[15]	2021	Machine Learning	71%	77%
[16]	2021	Machine Learning	68%	71%
[17]	2019	Machine Learning	-	75%

Paper provided a comprehensive review of the topic stating that AI models trained on clinicopathological features produce AUC results that range from 74% to 77%.

III. Materials and Methods

A. Data Source

We used an anonymized dataset from Baheya Foundation consisting of 950 medical records that are presented in a tabular format. Dataset contains 25 features and a binary ground truth (positive or negative). It combines patient data from radiology reports, clinical data, and pathological data.

B. Data Preparation

We began by exploring the data to analyze the dataset. Our dataset contained numerical and categorical features as well as missing values and outliers. Categorical data has three types: nominal, ordinal, and binary. We assessed the data balance and identified an imbalance, with 75% of instances belonging to the positive class and 25% to the negative class. Additionally, We examined the cardinality and analyzed the distribution of both numerical and categorical features. We determined that a large percentage of our data was missing. Careful preprocessing was applied to ensure the quality and relevance of the data used for training models, which is important to achieve high accuracy. We started by removing duplicate records, correcting spelling errors or inconsistencies,and eliminating outliers. Categorical data needs to be converted to numerical by encoding. Ordinal encoding was applied to ordinal data. Nominal and Binary data were encoded using target encoding, which led to the best results as it preserves the number of columns which handles multicollinearity. Categorical features were then imputed using MICE Imputation to fill their missing values. Numerical features were scaled then imputed using KNN Imputer to fill their missing values. SMOTETomek, which is a variation of SMOTE technique, was used to handle class imbalance. We conducted a correlation analysis between the features and the target variable as part of our feature selection process. Table II below displays the selected features and their data type:

TABLE II
FEATURE DESCRIPTIONS OF THE DATASET

Numerical	Nominal	Ordinal	Binary
First_BMI	Other_Diseases	T	VTE
Age	Family_History	N	Menopausal_State
KI67	HER2	Grade	-
Tumor_Size	Site	-	LVI
-	TumorType	-	Laterality
-	-	-	ER
-	-	-	PR
-	-	-	Unilateral_Bilateral

C. Machine Learning Algorithms

To develop our prediction model, we employed different machine learning algorithms. Catboost, XGBoost, and Random Forest were our top three models, with Catboost being the best performer. Catboost is a gradient boosting technique. It is an ensemble learning method that merges weaker decision trees to build a powerful predictive model. It works by iteratively adding new models to the ensemble, each one trained to correct the errors made by the previous models. The algorithm begins with an initial estimate, usually the mean of the target variable. It then incrementally builds a collection of decision trees, where each tree seeks to minimize the errors from the preceding trees. The algorithm progressively builds the ensemble of trees by minimizing the loss function with gradient descent. At every stage, it calculates the negative gradient of the loss function from the current predictions and then trains a new tree to match this negative gradient. Mathematically, CatBoost can be represented as follows:

Given a training dataset with N samples and M features, where each sample is represented as (x_i, y_i) with x_i being a vector of M features and y_i as the corresponding target variable, CatBoost strives to learn a function $F(x)$ that predicts the target variable y.

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

where:

- $F(\mathrm{x})$ is the overall prediction function.
- $F_0(\mathrm{x})$ is the initial prediction function.
- M is the number of trees in the model.
- n is the number of samples in the training dataset.
- $\sum_{m=1}^{M}$ is the summation over all M trees.
- $\sum_{i=1}^{N}$ is the summation over all N samples in the training dataset.
- $f_m(x_i)$ is the prediction of the m-th tree

We employed hyperparameter tuning for the CatBoost model to optimize its performance.The CatBoost model is configured with the following parameters:

- classifier_learning_rate: Learning rate of Classifier

- classifier_depth: Maximum depth of the trees
- classifier_depth: Number of boosting stages

A 10-fold cross-validation technique was adopted to ensure a reliable assessment of the CatBoost model's generalizability to unseen data.

D. Website

1. User Roles and Responsibilities:

Our system has several users, each with its own roles and responsibilities.

Admin has a panel dedicated to monitoring the system server status, CPU and memory usage, and the sending and receiving of packages. Additionally, the Admin manages the routing of SSL certificates using NGINX and Certbot. The admin can view the patients' records that have been entered and assessed by the doctors.

Head Doctor has the capabilities to sign up other doctors, and view all doctors. The Head Doctor can filter patients' based on features for instance they can get records by tumour type. Add to that, they can also view the number of entered patients' assessments for prediction in our system. The Head Doctor can showcase all patients' data, such as charts and figures, that illustrate the distribution frequency of each medical feature value.

A Doctor's role is to insert a patient's medical features to receive predictions about the likelihood of metastasis. Furthermore, they can view all patients and generate a variety of plots and figures using medical features.

Another prominent role is the data analyst where they have the capability to view a feature statistics dashboard, enabling analysis such as correlating features with each other and with the target variable and display the evaluation metrics like the ROC Curve and accuracy. Moreover, they can visualise feature distributions using various charts.

2. Security Measures:

Since our system contains medical data, it was crucial for us to implement security measures to our system. First for authentication and authorization we use tokens to verify the identity of our users and limit accessibility. Add to that, only admin is to sign up head doctors and data analysts and head of doctors are authorised to sign up for doctors. In addition all the generated passwords are saved encrypted and it can be changed by users later on. Another prominent feature is that in order to prevent automated abuse like Brute Force Attack, CAPTCHA is embedded in our sign in feature. Furthermore, our server is configured to use the SSL certificate for secure connections.

3. Usability Features:

We have prioritized usability features in our website design to ensure a smooth and efficient user experience. The whole website has simple forms and clear labels make it easy for new users to learn how to use the app. A consistent and logical layout like the physician patients management view where the sidebar and titles use consistent terminology and layout that helps returning users remember how to navigate the app. Visualisations like interactive bar charts and histograms like Tumour Size Distribution, help doctors explore their data more effectively. Interactivity as well was highly considered, dynamic visualisation where the dashboard dynamically generates interactive graphs and charts based on the filtered data.Pop up windows like the one providing the prediction to the doctor was implemented to grab attention.

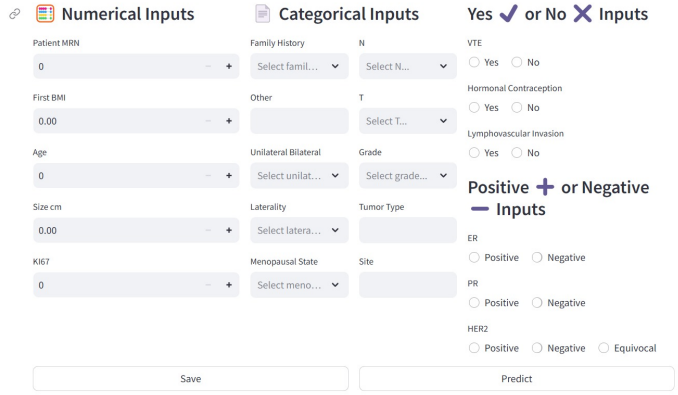

Fig. 1. Breast Cancer Metastasis Prediction Page

IV. Overall Experimental Workflow

After finishing the data collection phase, we performed data exploration to analyse the data. Subsequently, we pre-pocessed the data. Pre-processing starts with data cleaning from inconsistent or wrong entries then it differes based on whether a feature is numerical or categorical. The following figure illustrates the pre-processing steps:

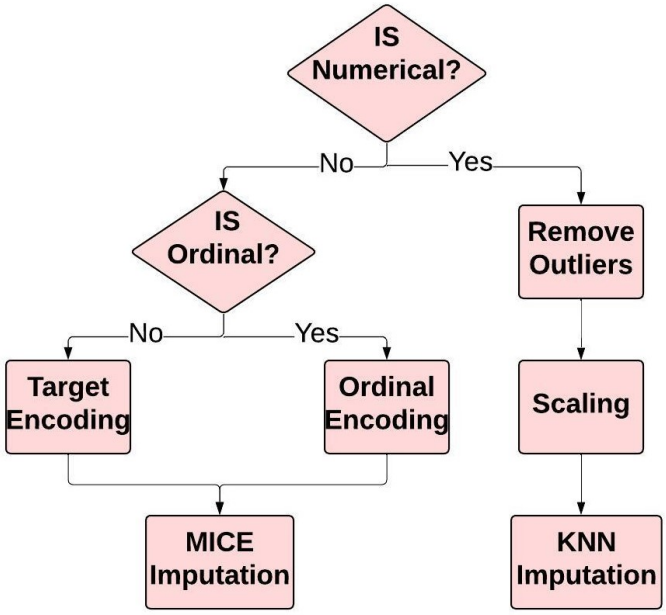

Fig. 2. Data Pre-Processing Steps

After data pre-processing, data imbalance issue is handled by SMOTETomek technique. Next, Feature Engineering is applied by calculating features' importance and the correlation of features. This leads to dropping extra features. At the end, the machine learning model will be trained in order to make predictions on unseen data.

To handle our workflow, we used Scikit-learn Pipeline. Scikit-learn Pipeline is a tool that combines multiple preprocessing steps with a machine learning model into a single object. This allows for seamless integration of data preprocessing and model fitting, simplifying the workflow and reducing the risk of data leakage.

V. Results

After careful pre-pre-processing of our data, we applied feature engineering concepts to select the features that will be trained by our model. 10-fold cross-validation technique was adopted to ensure a reliable assessment of the model's generalizability to unseen data. Catboost model had the best performance from all the tried models. Notably, the CatBoost model achieved an accuracy of 83% an F1 Score of 82%, and an AUC of 80%, demonstrating its effectiveness in prediction. The following graph shows the ROC Curve of the 10 folds and the mean AUC Value:

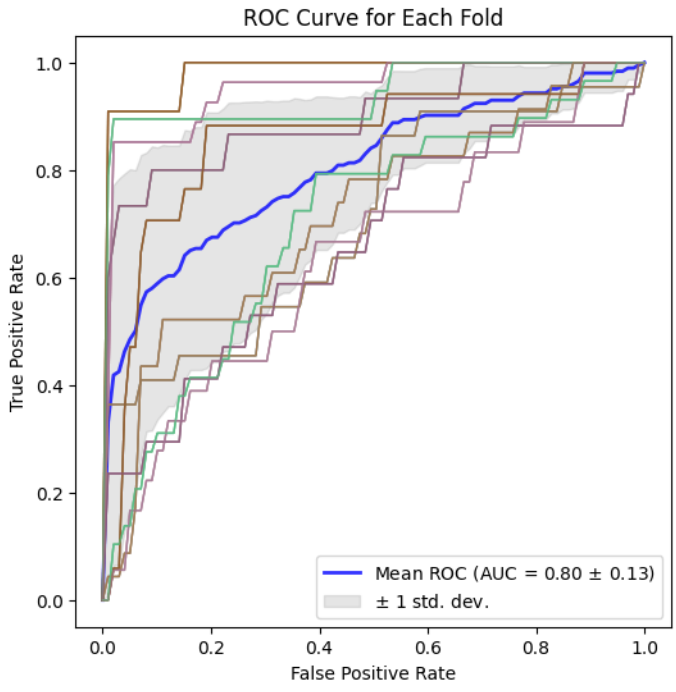

Fig. 3. Receiver operating characteristics (ROC) of 10 folds

VI. Conclusion

In this study, our machine learning model demonstrated the ability to accurately predict the involvement of axillary lymph node metastasis. In conclusion, our system provides a reliable tool that can assist healthcare professionals to predict whether patients have metastasis. Our system empowers doctors with a comprehensive toolset for managing patient records. This includes functionalities for data entry, record filtering, and advanced visualization tools.

VII. Future Work

For future work, we aim to connect our system to a platform that allows the data analyst to develop a code that can perform data analysis tasks that are specified by the analyst according to his desire. In other words, we aim that our system becomes able to adapt to the specific needs of each data analyst who uses it.

References

[1] https://www.wcrf.org/cancer-trends/breast-cancer-statistics/

[2] https://www.unitypoint.org/news-and-articles/most-dangerous-cancers-in-men-and-women

[3] https://www.nationalbreastcancer.org/breast-cancer-facts/

[4] https://gco.iarc.who.int/media/globocan/factsheets/populations/818-egypt-fact-sheet.pdf

[5] https://www.nationalbreastcancer.org/breast-cancer-facts/

[6] N. U. Dogan, S. Dogan, G. Favero, C. Köhler, and P. Dursun, "The Basics of Sentinel Lymph Node Biopsy: Anatomical and Pathophysiological Considerations and Clinical Aspects," J. Oncol., vol. 2019, no. 3415630, Jul. 2019. [Online]. Available: https://doi.org/10.1155/2019/3415630. PMCID: PMC6699370. PMID: 31467535.

[7] G. Qiao, Y. Cong, H. Zou, J. Lin, X. Wang, X. Li, Y. Li, and S. Zhu, "False-negative Frozen Section of Sentinel Lymph Node Biopsy in a Chinese Population with Breast Cancer," Anticancer Res., vol. 36, no. 3, pp. 1331–1337, Mar. 2016.

[8] "False Negative Rate of Sentinel Lymph Node Biopsy on Intraoperative Frozen Section in Early Breast Cancer Patients: An Institutional Experience," Original Article, vol. 13, pp. 312–315, Oct. 2021.

[9] S. M. Mokhtar, O. Mahmoud, R. Wessam, and E. Khallaf, "In early-stage breast cancer, sentinel lymph node biopsy can save unnecessary axillary dissection compared with fine-needle aspiration cytology for indeterminate axillary lymph nodes," Egypt. J. Surg., vol. 40, no. 1, pp. 209–215, Jan.–Mar. 2021. DOI: 10.4103/ejs.ejs_292_20.

[10] https://www.cancer.gov/news-events/cancer-currents-blog/2017/breast-cancer-lymph-node-removal

[11] H. Verbelen, W. Tjalma, J. Meirte, and N. Gebruers, "Long-term morbidity after a negative sentinel node in breast cancer patients," Eur. J. Cancer Care, first published May 3, 2019. DOI: 10.1111/ecc.13077.

[12] R. Shahriarirad, S. M. M. Yazd, R. Fathian, M. Fallahi, Z. Ghadiani, and N. Nafissi, "Prediction of sentinel lymph node metastasis in breast cancer patients based on preoperative features: a deep machine learning approach," Sci. Rep., vol. 14, Art. no. 1351, 2024.

[13] J. Vrdoljak et al., "Applying Explainable Machine Learning Models for Detection of Breast Cancer Lymph Node Metastasis in Patients Eligible for Neoadjuvant Treatment," Cancers (Basel), vol. 15, no. 3, p. 634, Feb. 2023. DOI: 10.3390/cancers15030634. PMCID: PMC9913601. PMID: 36765592.

[14] C. Jiang et al., "Prediction of lymph node metastasis in patients with breast invasive micropapillary carcinoma based on machine learning and SHapley Additive exPlanations framework," Front Oncol., vol. 12, p. 981059, Sep. 2022. DOI: 10.3389/fonc.2022.981059. PMCID: PMC9520536. PMID: 36185290.

[15] L. Meng et al., "Development of a prediction model based on LASSO regression to evaluate the risk of non-sentinel lymph node metastasis in Chinese breast cancer patients with 1–2 positive sentinel lymph nodes," Sci. Rep., vol. 11, p. 19972, Oct. 2021. DOI: 10.1038/s41598-021-99522-3. PMCID: PMC8497590. PMID: 34620978.

[16] A. Fanizzi et al., "Sentinel Lymph Node Metastasis on Clinically Negative Patients: Preliminary Results of a Machine Learning Model Based on Histopathological Features," Appl. Sci., vol. 11, no. 21, p. 10372, Nov. 2021. DOI: 10.3390/app112110372.

[17] Y.-J. Tseng et al., "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies," PMID: 31103449. DOI: 10.1016/j.ijmedinf.2019.05.003.