

Machine Learning–Based Multioutput Prediction of Surgery, Radiotherapy, and Chemotherapy in Breast Cancer

Lukas Weber^{1*}, Sophie Janssen², Jonas Richter¹, Elena Fischer³

¹ Department of Breast Oncology and ML-Based Treatment Planning, Faculty of Medicine, University of Stuttgart, Stuttgart, Germany.

² Department of Multioutput Prediction and Clinical Decision Support, Faculty of Medicine, Eindhoven University of Technology, Eindhoven, Netherlands.

³ Department of Breast Cancer Multimodality Therapy, Faculty of Medicine, University of Basel, Basel, Switzerland.

*E-mail ✉ lukas.weber@uni-stuttgart.de

Abstract

Breast cancer represents the leading malignancy across the globe and is characterized by a continuous rise in both incidence and mortality rates each year. This study introduces an innovative predictive framework for determining appropriate breast cancer interventions—surgery, radiotherapy, and chemotherapy—using a multi-output classifier that leverages algorithms such as CatBoost, XGBoost, NN, and NN Binary. To address the urgent requirement for reliable therapeutic guidance, the developed model aims to substantially improve the accuracy of treatment outcome predictions. It performs exceptionally well in estimating surgical interventions, particularly with Neural Networks (both NN and NN Binary), delivering 97% accuracy and 98% F1-scores. Radiotherapy predictions yield a moderate overall accuracy of around 63%, with strong recall reaching up to 84%. For chemotherapy, the model maintains stable accuracy and precision levels of 82%, with AUC-ROC scores as high as 89%, reflecting superior classification capability. By merging multi-output classification techniques with state-of-the-art algorithms, this work aims to generate more patient-specific treatment forecasting tools, supporting the shift toward personalized oncology strategies and meeting the growing expectations of precision medicine in cancer management.

Keywords: Breast cancer, Machine learning, Treatment prediction, SEER dataset, Computer-aided treatment prediction (CATP), Clinical decision support system

Introduction

Breast cancer arises when breast cells proliferate uncontrollably. Multiple subtypes exist, classified by the specific cells that have undergone malignant changes. Representing 12.5% of all newly reported cancer diagnoses each year, breast cancer constitutes the most widespread cancer type internationally. In 2020, it caused

685,000 fatalities among 2.3 million women diagnosed with cancer. At the conclusion of that year, 7.8 million women had been diagnosed with breast cancer over the prior five-year period. Estimates for 2023 anticipate 297,790 new cases of invasive breast cancer among women in the United States, confirming its position as the predominant global malignancy [1, 2]. Clinically, malignant growths are regarded as positive, in contrast to benign growths, which are viewed as negative. Each category contains distinct subgroups requiring individualized recognition because they often differ in prognosis and therapeutic requirements. Reliable separation of these subgroups is critical for selecting appropriate interventions [3-8].

Access this article online

<https://smerpub.com/>

Received: 27 August 2025; Accepted: 19 December 2025

Copyright CC BY-NC-SA 4.0

How to cite this article: Weber L, Janssen S, Richter J, Fischer E. Machine Learning–Based Multi-Output Prediction of Surgery, Radiotherapy, and Chemotherapy in Breast Cancer. J Med Sci Interdiscip Res. 2025;5(2):151-62. <https://doi.org/10.51847/JWCyHkpkc7>

Treatment strategies for cancer fall into two main categories—local and systemic—guided by the disease’s nature and progression stage. Local modalities, including surgery and radiation, focus exclusively on the tumor site without affecting other bodily regions. Systemic therapies, however, deploy pharmaceutical agents to target cancer cells body-wide. These drugs can be taken orally or introduced into the bloodstream and encompass chemotherapy, hormone therapy, targeted agents, and immunotherapy. Administration may occur before surgery (neoadjuvant), after surgery (adjuvant), or at both stages [9-11]. In practice, most individuals undergo combined local control measures together with systemic treatment to manage any distant disease spread.

Machine Learning (ML) is a key branch of Artificial Intelligence (AI) that enables computational systems to detect patterns and learn from data without direct programming. Its primary categories include supervised learning, which trains on labeled datasets to forecast or categorize outcomes; unsupervised learning, which explores unlabeled data for underlying structures; and reinforcement learning, which optimizes actions according to reward signals. Widely adopted supervised techniques comprise Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Random Forest (RF), Linear Discriminant Analysis (LDA), and Logistic Regression (LR) classifiers [12]. ML frameworks can process radiomic features derived from imaging as inputs. By integrating various radiomic elements, such as tumor profiles, into unified scores, these algorithms help estimate disease likelihood and guide optimal treatment decisions [13]. Traditional ML approaches like LR or Decision Trees (DT) display shortcomings in their predictive processes. Operating strictly within the initial feature space, they rarely construct intermediate representations despite elaborate feature engineering. Such models succeed primarily when input variables carry strong predictive power, a situation not universally present in complex medical datasets [14].

Deep Learning (DL), an extension of ML, utilizes multi-layered neural architectures to uncover sophisticated patterns straight from unprocessed inputs. Neural Networks (NN) are typically developed through supervised training, pairing each example with its designated target. DL’s success in resolving diagnostic issues has fueled substantial interest in AI-driven healthcare solutions [15].

Oncology care involves intricate coordination across multiple specialties and therapeutic options.

Multimodality regimens have gained prominence owing to refinements in operative procedures, chemotherapeutic agents, digital technologies, and precision molecular and radiotherapeutic innovations. However, fragmented communication among specialists can impede timely treatment design and implementation [16]. Breast cancer specialists contend with an immense and ever-expanding knowledge base. Establishing large-scale genomic and clinical repositories along with analytical computational tools could accelerate therapeutic progress and expedite guideline revisions. These challenges are intensified by restricted time for information documentation and retrieval during actual patient encounters [17]. Research surveying 1117 oncologists found that they allocate roughly 4.6 hours per week to remaining updated in their field, while devoting 53 hours to clinical duties and administrative responsibilities [18].

Clinical Decision Support (CDS) systems consist of computational applications intended to alleviate the intellectual demands created by this information surge. Many derive from foundational AI and expert system initiatives [19]. Within healthcare, AI tools excel at extracting insights from unstructured sources like textual records via natural language processing or from expansive structured repositories through ML methodologies. They leverage reasoning engines to contextualize knowledge, formulate hypotheses, and generate physician-actionable assessments after appropriate curation [20]. Advanced CDS platforms go further by supporting individualized, evidence-grounded choices that surpass conventional database queries and Electronic Patient Records (EPRs). Effective systems can continuously evaluate comprehensive patient information and deliver pertinent notifications, prompts, and structured guidance [21].

OncoDoc functions as a decision-tree-based knowledge navigation instrument that helps physicians contextualize patient variables to construct accurate case representations. As a digital screening aid, it supplies either guideline-supported treatment pathways or suitable clinical trial options by aligning patient profiles with established protocols or trial criteria [22]. MATE (Multidisciplinary Team Assistant and Treatments Elector) represents another CDS platform tailored for breast cancer Multi-Disciplinary Team (MDT) discussions; it was assessed at London’s Royal Free Hospital to advance patient-focused, evidence-supported choices [21]. IBM Watson for Oncology (WFO),

developed in collaboration with Memorial Sloan Kettering (MSK), serves as a CDS resource to equip medical oncologists with specialized recommendations derived from essential disease and patient details. WFO addresses nuances absent from standard guidelines by incorporating logic from MSK expertise and machine learning on authentic and simulated cases. Factors such as breast surgery, radiotherapy, clinical genetics, and fertility preservation figure prominently in decisions regarding adjuvant systemic therapy for early-stage disease [23].

Although these decision-support technologies continue to mature, their integration into everyday clinical workflows remains limited. Many prior evaluations relied on restricted or institution-specific datasets, constraining external validity and practical deployment. Performance has proven inconsistent across varying cancer stages [17, 23]. Further investigation is necessary to facilitate wider implementation and strengthen clinical utility.

Earlier research outlined a framework for a Computer-Aided Treatment Prediction (CATP) System targeting breast cancer. It employed multi-class classification across four prominent algorithms—Decision Trees, Random Forest, XGBoost, and CatBoost—securing a maximum overall AUC of 0.91 with the Random Forest model [24]. The present investigation extends this research by exploring a multi-output classifier paradigm. Its primary goal is to construct a model capable of markedly improving the accuracy of predictions related

to breast cancer treatment results. This approach is expected to improve the system's adaptability to heterogeneous treatment responses and distinct patient characteristics.

Materials and methods

Dataset collection

The study was initiated by retrieving the original dataset from the November 2020 edition of the SEER Research Plus database, which includes patient information collected between 1992 and 2018 [25]. A careful selection procedure identified 25 clinically relevant attributes, producing 1,048,575 breast cancer (BC) entries. Because of inconsistencies in data recording practices across different SEER releases, we performed extensive preprocessing and validation steps to ensure reliable data quality and uniformity. Following feature selection and intensive cleaning, the processed dataset contained 335,990 entries featuring nineteen attributes (**Table 1**). These attributes include patient demographic details such as age brackets, tumor-specific properties, indicators of disease spread, and three variables denoting the primary treatment options of surgery, radiotherapy, and chemotherapy. Since the analysis focuses exclusively on whether each treatment is advised, the treatment fields were recoded into binary values (0/1). These binary labels subsequently served as the ground truth for training the multi-output classifier.

Table 1. Attributes obtained from the SEER database.

Feature	Paraphrased description
Age	Patient's age at the time of breast cancer diagnosis.
Laterality	Specifies whether the primary breast tumor developed in the left or right breast.
Histology (ICD-O-3)	ICD-O-3 code defining the microscopic characteristics of tumor cells and tissues. Histological classification plays a key role in cancer staging, treatment selection, prognosis, and disease progression.
Breast subtype	Breast cancer subtype generated by integrating information from ER Status Recode, PR Status Recode, and Derived HER2 Recode.
Tumour size	Records the measured size of the primary tumor.
Lymph nodes	Indicates the status of lymph node involvement.
Regional nodes evaluated	Represents the total count of regional lymph nodes surgically removed and assessed by a pathologist.
Regional nodes positive	Specifies the exact number of examined regional lymph nodes confirmed to contain metastatic disease.
Metastasis at distant lymph nodes	Indicates whether metastasis is present in distant lymph nodes.
Stage	Overall clinical stage assigned to the cancer.
T	AJCC T classification describing the extent and size of the primary tumor.

N	AJCC N classification indicating the degree of spread to regional lymph nodes.
M	AJCC M classification identifying whether distant metastatic spread has occurred.
ER	Indicates the presence or absence of estrogen receptor (ER) expression in the tumor.
PR	Indicates the presence or absence of progesterone receptor (PR) expression in the tumor.
HER2	Indicates whether the tumor expresses the human epidermal growth factor receptor 2 (HER2).
Surgery	Specifies whether surgical treatment is recommended.
Radiotherapy	Specifies whether radiation therapy is recommended.
Chemotherapy	Specifies whether chemotherapy is recommended.

Multioutput classifier

A defining feature of the multi-output classifier technique is that it creates a dedicated classifier for each target variable. This design supports detailed and simultaneous prediction of several related outcomes. This structure is especially valuable in clinical decision-making contexts, where multiple treatment options must be evaluated together with high granularity. Functionally, separate models are trained for every output. For unseen cases, each model produces an independent prediction, which is then aggregated into a single unified output vector representing the full set of recommended interventions. This strategy offers considerable adaptability for multifaceted multi-label problems. Overall, it enhances standard single-output classifiers by enabling them to address scenarios with multiple interdependent targets [26] proficiently.

In this work, we present a new predictive framework for breast cancer interventions. It employs multi-output classifier methods using CatBoost, XGBoost, NN, and NN Binary algorithms. The primary intention is to construct a system capable of achieving markedly higher accuracy in forecasting appropriate breast cancer treatments. This design is projected to accommodate better the complex interplay of patient-specific factors and varying therapeutic responses. **Figure 1** depicts the overall process and logic of the proposed model.

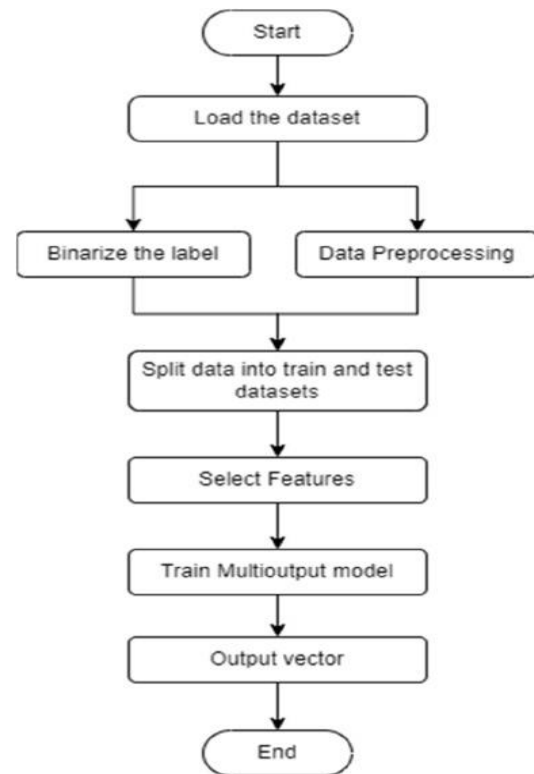


Figure 1. Overview of the multi-output model.

Standard evaluation metrics were applied to compare model effectiveness and determine the optimal solution for anticipating breast cancer treatment needs. Based on these criteria, the best-performing algorithm was selected as the final model, improving personalization and reliability in treatment recommendations for breast cancer patients.

Feature selection

Feature selection aims to isolate a relevant group of input variables that adequately represent the underlying data while minimizing irrelevant information and noise, without compromising predictive capability. This process has emerged as a robust preparatory step in machine learning workflows, especially for high-dimensional datasets [27-29].

Following the approach outlined in an earlier investigation [24], the present study employs the Shapley Additive Explanations (SHAP) technique, grounded in the Shapley value principle from cooperative game theory [30], alongside Shap-hypetune. The latter is a specialized library that integrates hyperparameter tuning and feature selection within a single workflow to identify the optimal feature count and parameter configuration [31]. For a given model $f(x_1, x_2, \dots, x_d)$, the features ranging from 1 to d can be viewed as participants in a cooperative game, where the payoff v reflects the contribution or importance of a particular subset. The Shapley value $\phi_v(i)$ then quantifies the marginal contribution of feature i to the overall outcome [32].

The SHAP value computation process in this research starts with dataset sampling, preserving a defined proportion of instances to balance computational efficiency with representational fidelity. A greedy selection strategy is adopted, emphasizing features that show strong, stable contributions through a percentile-based threshold. The SHAP Explainer (KernelSHAP or TreeSHAP) is used to compute the mean absolute SHAP value for each feature, evaluating its global importance regardless of effect direction. A consistency metric, defined as the standard deviation relative to the mean SHAP value, is used to retain features with high impact and low variability. Features close to the threshold that maintain stable performance are also incorporated. The ultimate feature set is refined by merging the selected variables and eliminating redundancies. This procedure aims to improve the predictive model's effectiveness and transparency by focusing on features that are both influential and dependable, as identified by SHAP analysis.

Model evaluation

The performance of the developed model will be rigorously assessed through a detailed comparative examination utilizing a broad range of established performance indicators, namely accuracy, precision, recall, F1-score, and AUC-ROC [33-38].

Accuracy is computed as the sum of True Positive (TP) and True Negative (TN) instances divided by the total of TP, TN, False Positive (FP), and False Negative (FN):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision measures the proportion of correctly identified positive predictions among all positive predictions made by the model and is expressed as a percentage. It reflects the degree of confidence one can place in the model's positive classifications:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall evaluates the model's ability to correctly identify all actual positive instances within the dataset, thereby indicating its effectiveness in detecting the positive class. It is determined by dividing the number of TP by the total number of actual positive cases:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

The F1-score is the harmonic mean of precision and recall, ranging from 0 to 1, where higher values indicate better performance [39]. This metric provides a balanced assessment of classifier effectiveness across classes and is particularly valuable for imbalanced datasets or when all categories are equally important. It is calculated as follows:

$$F1 - \text{Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

The Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) corresponds to the area beneath the ROC curve and yields values between 0 and 1.0. Since random classification produces a diagonal line with an area of 0.5, effective classifiers are expected to achieve an AUC exceeding 0.5. A key statistical advantage of the AUC is that it equals the probability that a randomly chosen positive instance will receive a higher score than a randomly chosen negative instance. For binary classification tasks, the AUC can be expressed using the formula:

$$\text{AUC} = \frac{S_p - \frac{n_p(n_p + 1)}{2}}{n_p n_n} \quad (5)$$

Where S_p denotes the sum of the ranks of all positive examples, and n_p and n_n indicate the numbers of positive and negative examples, respectively. Empirical and theoretical evidence supports the superiority of the AUC metric over accuracy for evaluating classifier

performance and selecting optimal models during training [40].

Results and Discussion

We performed a comprehensive evaluation to assess the effectiveness of the proposed models. We systematically optimized hyperparameters for the CatBoost, XGBoost, NN, and NN Binary configurations. We applied GridSearchCV to CatBoost and XGBoost to explore optimal combinations of key parameters. For CatBoost, the tuned parameters included learning rate (controlling convergence speed), depth (governing model complexity), and l2_leaf_reg (a regularization term to mitigate overfitting). The optimal settings identified were 0.03 for learning_rate, 6 for depth, and 3 for l2_leaf_reg. For XGBoost, we optimized max_depth, min_child_weight, and gamma to balance complexity and generalization. The best values obtained were 4 for max_depth, 6 for min_child_weight, and 0 for gamma. We optimized the Neural Network and NN Binary

models using Optuna, an automated hyperparameter search framework suitable for TensorFlow/Keras architectures [41]. Optimization targeted network structure elements such as layer count, activation functions, optimizer selection, learning rate, and dropout probabilities to improve generalization while controlling overfitting. The optimal configuration determined by Optuna consisted of n_layers = 1, dropout_rate = 0.48, learning_rate = 0.00335, n_units_l0 = 55, and n_units_l1 = 95. All models underwent five-fold cross-validation to verify stability across varied data partitions.

Additionally, we reserved 20% of the dataset as an independent validation set to ensure unbiased performance estimation. We computed evaluation metrics—including accuracy, precision, recall, F1-score, and ROC-AUC—separately for each treatment category (surgery, radiotherapy, and chemotherapy). These indicators provide essential insights into the models' capacity to classify positive and negative cases correctly, their overall reliability, and their ability to distinguish between classes (**Table 2**).

Table 2. Performance metrics for each treatment type across the evaluated models.

Outcome	Evaluation measure	CatBoost	XGBoost	Neural network (NN)	Binary neural network
Surgery	Accuracy	97	97	96	97
	Precision	88	87	97	97
	Recall	80	79	99	99
	F1 Score	83	82	98	98
	AUC-ROC	80	79	95	97
Radiotherapy	Accuracy	63	62	61	61
	Precision	63	62	56	58
	Recall	63	62	84	73
	F1 Score	62	62	67	65
	AUC-ROC	63	62	65	66
Chemotherapy	Accuracy	82	82	82	82
	Precision	82	82	77	78
	Recall	82	81	77	77
	F1 Score	82	82	77	77
	AUC-ROC	82	82	89	89

Figure 2 illustrates the accuracy of the various models on the SEER dataset. The models performed exceptionally well in predicting surgical interventions. Both XGBoost, CatBoost, and NN Binary attained an outstanding accuracy of 0.97, while the standard NN model reached 0.96. These outcomes indicate strong predictive capability in the surgical domain. In contrast, performance declined considerably for radiotherapy predictions. Accuracy fell to 0.63 for CatBoost and 0.62 for XGBoost, with both NN and NN Binary models

recording 0.61. This reduction suggests that forecasting radiotherapy recommendations is more challenging, potentially because of more complex underlying variables or subtle patterns in the data. All models achieved a consistent accuracy of 0.82 for chemotherapy predictions. This stability implies that the dataset contains clear indicators for chemotherapy decisions or that these factors are comparatively straightforward to model. Given chemotherapy's established role in treatment protocols, the models appear well suited to

predicting its use. Overall, while surgery and chemotherapy predictions maintained high accuracy, radiotherapy forecasting showed noticeably weaker results across all algorithms. This disparity highlights inherent complexities in modeling radiotherapy outcomes and points to opportunities for future refinement in feature engineering or model architecture.

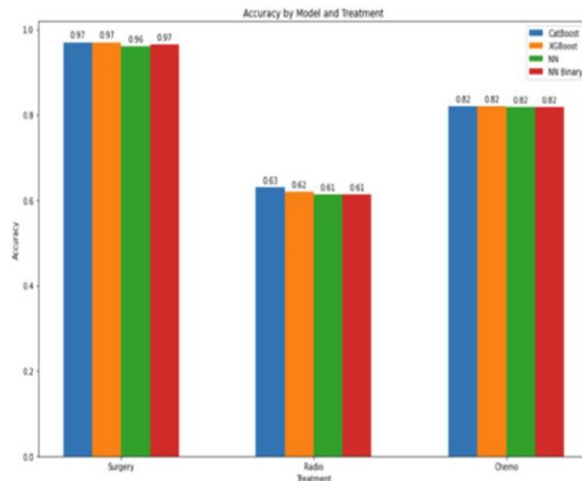


Figure 2. Accuracy performance by model and treatment modality.

Figure 3 displays the precision attained by the four models across the three treatment categories in a bar chart format. Precision reflects the models' reliability in identifying positive cases without generating excessive false positives. For surgery, the NN models achieved the highest precision (0.97), demonstrating strong performance in correctly classifying surgical recommendations. CatBoost and XGBoost followed closely with precision scores of 0.88 and 0.87, respectively. Precision decreased markedly for radiotherapy across all models. CatBoost recorded 0.63, slightly outperforming XGBoost at 0.62. The standard NN model showed a precision of 0.56, while NN Binary improved marginally to 0.58. For chemotherapy, CatBoost and XGBoost both maintained a strong precision of 0.82. The NN model registered 0.77, and NN Binary reached 0.78. In summary, the NN models excelled in surgical precision, while CatBoost and XGBoost offered greater consistency across radiotherapy and chemotherapy predictions.

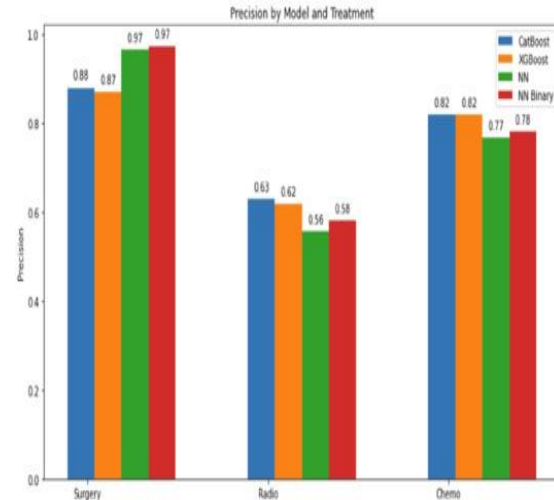


Figure 3. Precision performance by model and treatment modality.

Figure 4 presents the recall values for the four models across the three treatment types. Recall is particularly important in medical applications because it indicates the models' ability to identify all relevant positive cases. Both NN and NN Binary models demonstrated excellent recall exceeding 0.99 for surgery, indicating a very low likelihood of missing true positive cases. CatBoost achieved a solid recall of 0.80, closely followed by XGBoost at 0.79. Recall declined across the board for radiotherapy predictions. The standard NN model attained the highest recall of 0.84 in this category, substantially outperforming CatBoost (0.63) and XGBoost (0.62). NN Binary recorded a recall of 0.73. For chemotherapy, CatBoost and XGBoost both achieved a recall of 0.81, while the NN and NN Binary models obtained 0.77. Collectively, the results show strong recall performance for surgery across all models, with NN variants performing best in radiotherapy. CatBoost and XGBoost held a slight advantage in chemotherapy recall. These findings can inform model selection depending on treatment type and the clinical priority of minimizing missed cases.

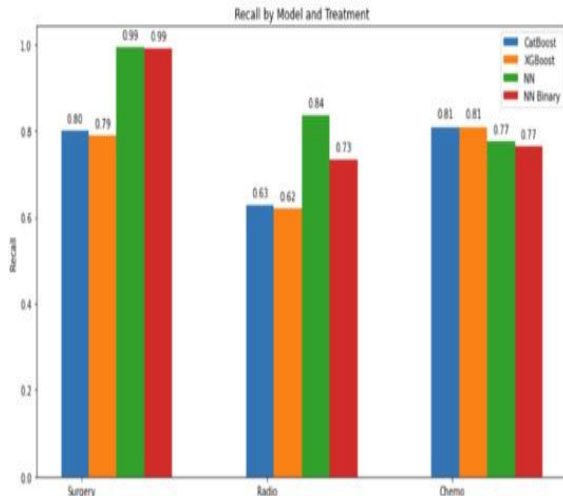


Figure 4. Recall performance by model and treatment modality.

Figure 5 illustrates the F1-scores achieved by the four models across the three treatment categories. In clinical contexts, where both false positives and false negatives carry significant consequences, the F1-score serves as a balanced indicator that integrates precision and recall. For surgical predictions, the NN and NN Binary models attained near-optimal F1-scores of 0.98, reflecting an excellent equilibrium between precision and recall. CatBoost followed with an F1-score of 0.83, and XGBoost with 0.82. These strong results highlight the models' effectiveness in surgical treatment forecasting, particularly their ability to identify true cases while limiting false predictions. Performance declined across all models for radiotherapy, likely reflecting the greater complexity and variability associated with this treatment modality. The NN model achieved an F1-score of 0.67, representing a reasonable yet reduced balance compared to its surgical performance. The NN Binary model followed with 0.65. Both CatBoost and XGBoost recorded an F1-score of 0.62, indicating potential for further refinement in balancing sensitivity and specificity. In the chemotherapy category, CatBoost and XGBoost both achieved an F1-score of 0.82, demonstrating consistent proficiency in balancing precision and recall. The NN and NN Binary models registered F1-scores of 0.77, which, while slightly lower, still reflect solid predictive capability. Overall, the NN models showed superior F1-scores for surgery, while performance differences narrowed for radiotherapy and chemotherapy, with CatBoost and XGBoost delivering competitive results. The NN Binary model's stable

performance across treatments suggests it offers dependable, well-balanced predictions, particularly in settings favoring simplified binary architectures. These observations can help select a model based on the specific requirements for balancing precision and recall in various therapeutic contexts.

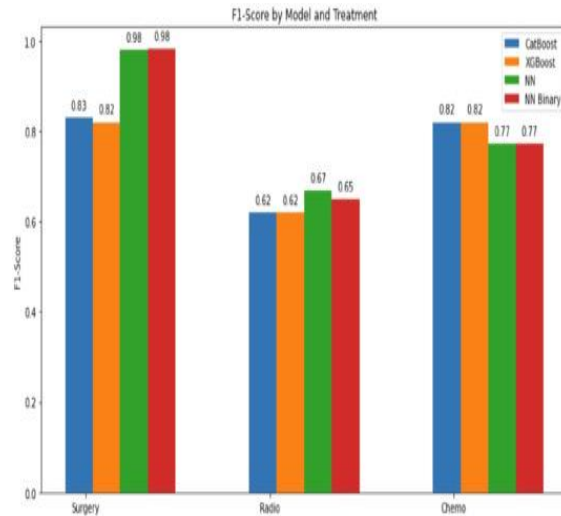


Figure 5. F1-score performance by model and treatment modality.

Figure 6 presents the AUC-ROC values for the four models across the three treatment scenarios in a bar graph. The AUC-ROC metric evaluates classification performance across varying decision thresholds. A value of 1.0 indicates perfect discrimination, whereas 0.5 corresponds to random classification with no discriminative power. In the surgical domain, the NN and NN Binary models achieved high AUC-ROC scores of 0.95 and 0.96, respectively, demonstrating an excellent ability to differentiate between patients requiring surgery and those who do not. CatBoost and XGBoost obtained scores of 0.79, indicating moderate but acceptable discriminative capacity. AUC-ROC values decreased for all models in radiotherapy predictions, underscoring the difficulty of distinguishing outcomes in this treatment area. The NN Binary model attained the highest score of 0.66, followed by the standard NN model at 0.65. CatBoost and XGBoost recorded 0.63 and 0.62, respectively. While all models faced challenges with radiotherapy, the neural network-based approaches performed marginally better. For chemotherapy, all models showed improved AUC-ROC scores. CatBoost and XGBoost achieved 0.82, reflecting good class separation. The NN and NN Binary models performed

best in this category, both reaching 0.89. This indicates a particular strength of neural network architectures in chemotherapy prediction tasks.

In summary, the results reveal robust discriminative performance for surgery and chemotherapy across the models, with the NN Binary model showing a clear advantage in surgical predictions. Radiotherapy prediction remains difficult for all algorithms, as evidenced by the lower AUC-ROC scores. The NN models, especially for surgery, generally outperformed CatBoost and XGBoost, suggesting their suitability when precise differentiation of patient outcomes is essential.

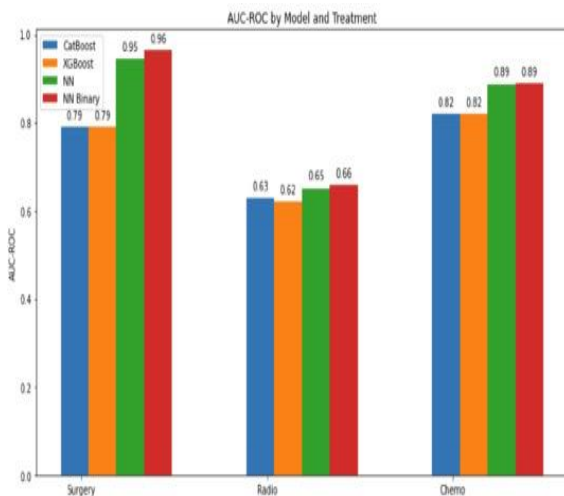


Figure 6. AUC-ROC performance by model and treatment modality.

Interpretability and clinical application

Successful integration of machine learning models into clinical settings requires both high accuracy and clear interpretability. Clinicians must understand the basis of predictions to incorporate model outputs into decision-making confidently. This research addresses interpretability through multiple complementary approaches. Feature importance is examined using SHAP values, which quantify each variable's contribution and highlight the primary drivers of predictions. Supporting visualization techniques, including SHAP summary plots, dependence plots, and force plots, enable clinicians to examine and interpret the role of individual features in specific predictions. Furthermore, SHAP values provide instance-level explanations that elucidate the reasoning for each prediction, an essential aspect in medical environments.

The model can also be incorporated into Electronic Patient Record (EPR) systems to deliver real-time predictions and accompanying explanations directly within routine clinical workflows, thereby facilitating immediate access for healthcare providers. Decision support platforms built on this model can assist clinicians by providing treatment recommendations with transparent rationales, enhancing understanding of the factors underlying each proposed intervention.

Conclusion

The overall effectiveness of the multi-output classifier models can be evaluated by examining their accuracy, precision, recall, F1-score, and AUC-ROC values for the three treatment categories: surgery, radiotherapy, and chemotherapy.

In surgical predictions, both the NN and NN Binary models demonstrated near-optimal precision and recall, resulting in strong AUC-ROC values and near-perfect F1-scores. CatBoost and XGBoost also exhibited excellent precision, recall, and AUC-ROC performance, confirming their high suitability for forecasting surgical interventions.

Performance metrics declined for all models in radiotherapy predictions, likely due to the greater complexity and variability of outcomes in this domain. Although AUC-ROC scores were generally lower across the four models, the NN model achieved the highest F1-score supported by strong precision and recall, along with an AUC-ROC value of 0.65. These findings indicate that radiotherapy treatment prediction remains challenging for all approaches, despite the NN model's relative advantage.

For chemotherapy outcome predictions, CatBoost and XGBoost delivered robust precision and recall scores. The NN-based models recorded slightly lower F1-scores but demonstrated a marginal advantage in AUC-ROC, suggesting enhanced ability to differentiate between patients who require chemotherapy and those who do not. In summary, CatBoost and XGBoost produced closely aligned results across all three treatment types. While highly effective for surgery and chemotherapy, both struggled with radiotherapy predictions. The standard and binary NN models excelled in surgery, performed comparably in chemotherapy, and outperformed traditional models in radiotherapy. The NN Binary variant offered particularly consistent performance throughout the treatment categories.

This detailed analysis indicates that the specific treatment context and the relative importance of individual performance metrics should guide model selection. For critical clinical decisions, prioritize models that provide the most favorable overall balance across metrics. In certain scenarios, the clinical cost of false positives versus false negatives may also influence choice. A notable limitation in comparative evaluations with prior work stems from the difficulty of replicating studies conducted on private or institutional datasets. By utilizing the publicly available SEER dataset, the current study promotes transparency and reproducibility, underscoring a key strength of the presented methodology.

Limitations and future work

The extensive evaluation of the multi-output model across multiple treatment prediction tasks reveals several limitations that merit further attention. First, the variation in performance between neural network models and conventional machine learning approaches such as CatBoost and XGBoost across different treatment categories underscores the difficulty of developing a universally robust predictive framework. The consistent decline in performance metrics for radiotherapy predictions across all models highlights a core issue: the inherent complexity and unpredictability of outcomes associated with this treatment modality. Existing modeling strategies may not adequately capture these characteristics.

Future investigations should explore techniques to optimize the recall-precision trade-off. Additionally, advancing modeling techniques to more effectively represent the subtleties of multifaceted treatment outcomes within multi-output frameworks is essential. This could be achieved by examining adaptive thresholding strategies or advanced ensemble techniques, which may preserve high recall while limiting precision losses, thereby enhancing overall model reliability. Hybrid architectures that integrate the respective strengths of traditional algorithms and neural networks represent another promising direction. Such developments may require adopting more advanced machine learning methods, using larger and more heterogeneous datasets for model training, integrating domain expertise, and including patient-specific clinical variables in the predictive process.

Subsequent research should also evaluate the clinical consequences of false positives and false negatives for

particular treatments. Furthermore, developing specialized performance metrics better aligned with therapeutic objectives would be beneficial. Ultimately, extending the research to encompass real-world clinical trial data and establishing iterative feedback mechanisms for continuous model refinement will substantially contribute to developing AI-powered decision support systems that are both accurate and practically valuable in clinical practice.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Breast cancer [Internet]. World Health Organization; 2021 [cited 2021 Oct 9]. Available from: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
2. Breast Cancer Facts and Statistics 2024 [Internet]. Breastcancer.org; 2024 [cited 2024 Mar 30]. Available from: <https://www.breastcancer.org/facts-statistics>
3. Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*. 1967;89(2):211-5. doi:10.1148/89.2.211
4. Dorrius MD, Jansen-Van Der Weide MC, Van Ooijen PMA, Pijnappel RM, Oudkerk M. Computer-aided detection in breast MRI: a systematic review and meta-analysis. *Eur Radiol*. 2011;21(8):1600-8. doi:10.1007/s00330-011-2091-9
5. Jalalian A, Mashohor SBT, Mahmud HR, Saripan MIB, Ramli ARB, Karasfi B. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clin Imag*. 2013;37(3):420-6. doi:10.1016/j.clinimag.2012.09.024
6. Friedewald SM, Rafferty EA, Rose SL, Durand MA, Plecha DM, Greenberg JS, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA*. 2014;311(24):2499-507. doi:10.1001/jama.2014.6095

7. Giger ML. Machine learning in medical imaging. *J Am Coll Radiol.* 2018;15(3):512-20. doi:10.1016/j.jacr.2017.12.028
8. Sheth D, Giger ML. Artificial intelligence in the interpretation of breast cancer on MRI. *J Magn Reson Imag.* 2020;51(5):1310-24. doi:10.1002/jmri.26878.
9. Waks AG, Winer EP. Breast cancer treatment: a review. *JAMA.* 2019;321(3):288-300. doi:10.1001/jama.2018.19323
10. Breast Cancer Treatment (Adult) (PDQ®)—Patient Version [Internet]. National Cancer Institute; [cited 2020 Oct 4]. Available from: https://www.cancer.gov/types/breast/patient/breast-treatment-pdq#_185
11. Breast cancer: symptoms, risk factors, diagnosis, treatment & prevention [Internet]. WebMD; [cited 2020 Aug 27]. Available from: <https://www.webmd.com/breast-cancer/understanding-breast-cancer-basics>
12. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman MLD, et al. Natural and artificial intelligence in neurosurgery: a systematic review. *Clin Neurosurg.* 2018;83(2):181-92. doi:10.1093/neuros/nyx384
13. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imag.* 2013;26(6):1045-57. doi:10.1007/s10278-013-9622-7
14. Gao Y, Gera KJ, Lewin AA, Moy L. New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. *AJR Am J Roentgenol.* 2019;212(2):300-7. doi:10.2214/AJR.18.20392
15. Codari M, Schiaffino S, Sardanelli F, Trimboli RM. Artificial intelligence for breast MRI in 2008–2018: a systematic mapping review. *Am J Roentgenol.* 2019;212(2):280-92. doi:10.2214/AJR.18.20389
16. Fennell ML, Das IP, Clauser S, Petrelli N, Salner A. The organization of multidisciplinary care teams: modelling internal and external influences on cancer care quality. *J Natl Cancer Inst Monogr.* 2010;40:72-80. doi:10.1093/jncimonographs/lgq010
17. Somashekhar SP, Sepúlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol.* 2018;29(2):418-23. doi:10.1093/ANNONC/MDX781
18. Shanafelt TD, Gradishar WJ, Kosty M, Satele D, Chew H, Horn L, et al. Burnout and career satisfaction among US oncologists. *J Clin Oncol.* 2014;32(7):678-86. doi:10.1200/JCO.2013.51.8480
19. Duda R, Shortliffe EH. Expert Systems Research. *Science.* 1983;220(4594):261-8.
20. Xu F, Sepúlveda MJ, Jiang Z, Wang H, Li J, Liu Z, et al. Artificial intelligence treatment decision support for complex breast cancer among oncologists with varying expertise. *JCO Clin Cancer Inform.* 2019;3:1-10. doi:10.1200/CCI.18.00117
21. Patkar V, Acosta D, Davidson T, Jones A, Fox J, Keshtgar M. Cancer multidisciplinary Team meetings: evidence, challenges, and the role of clinical decision support technology. *Int J Breast Cancer.* 2011;2011:831605. doi:10.4061/2011/831605
22. Séroussi B, Bouaud J. Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. *Artif Intell Med.* 2003;29(1-2):153-67. doi:10.1016/S0933-3657(03)00040-X
23. Seidman AD, Hudis CA, Burstein HJ, Mayer IA, McArthur HL, Sledge GW Jr, et al. Integration of multimodality treatment planning for early-stage breast cancer (BC) into Watson for Oncology, a Decision Support System: seeing the forest and the trees. *J Clin Oncol.* 2015;33(15_suppl):e12042. doi:10.1200/jco.2015.33.15_suppl.e12042
24. Abd Al Rahman E, Intan Raihana Ruhaiyem N, Bouchahma M, Imran Musa K, Author C. Framework for a computer-aided treatment prediction (CATP) system for breast cancer. *Intelligent Automation & Soft Computing.* 2023;36(3):3007-28. doi:10.32604/iasc.2023.032580
25. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1992–2018) [Internet]. Bethesda (MD): National Cancer Institute, DCCPS, Surveillance Research Plus Program, Surveillance Systems Branch; released April 2020, based on November 2020 submission.
26. Tsoumakas G, Katakis I. Multi-label classification. *Int J Data Warehous Min.* 2007;3(3):1-13. doi:10.4018/jdwm.2007070101

27. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: a data perspective. *ACM Comput Surv.* 2017;50(6):1-45. doi:10.1145/3136625
28. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. *Neurocomputing.* 2018;300:70-9. doi:10.1016/j.neucom.2017.11.077
29. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng.* 2014;40(1):16-28. doi:10.1016/j.compeleceng.2013.11.024
30. Young BH. Monotonic solutions of cooperative games. *Int J Game Theor.* 1985;14(2):65-72.
31. shap-hypetune · PyPI [Internet]. [cited 2022 Mar 11]. Available from: <https://pypi.org/project/shap-hypetune/>
32. Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler SA. Problems with Shapley-value-based explanations as feature importance measures. In: *Proceedings of the 37th International Conference on Machine Learning (PMLR)*; 2020; p. 5491-500.
33. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Cham: Springer; 2018. doi:10.1007/978-3-319-98074-4
34. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recogn Lett.* 2009;30(1):27-38. doi:10.1016/J.PATREC.2008.08.010
35. Mahmudah KR, Indriani F, Takemori-Sakai Y, Iwata Y, Wada T, Satou K. Classification of imbalanced data represented as binary features. *Appl Sci.* 2021;11(17):7825. doi:10.3390/app11177825
36. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng.* 2009;21(9):1263-84. doi:10.1109/TKDE.2008.239
37. Mohammed H, Md Naser S. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Management Process.* 2015;5(2):1-11. doi:10.5121/ijdkp.2015.5201
38. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36.
39. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol.* 2011;2(1):37-63.
40. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng.* 2005;17(3):299-310.
41. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; 2019 Aug 4-8; Anchorage (AK). New York: ACM; 2019. doi:10.1145/3292500.3330701