# Advancing patient care: Machine learning models for predicting grade 3+ toxicities in gynecologic cancer patients treated with HDR brachytherapy

Andres Portocarrero-Bonifaz[1,2*]; Salman Syed[1]; Maxwell Kassel[1]; Grant W. McKenzie[1]; Vishwa M. Shah[3]; Bryce M. Forry[1]; Jeremy T. Gaskins[4]; Keith T. Sowards[1]; Thulasi Babitha Avula[1]; Adrianna Masters[1]; Jose G. Schneider[5]; Scott R. Silva[1]

[1]Department of Radiation Oncology, Brown Cancer Center, University of Louisville School of Medicine, Louisville, KY, United States of America

[2]Department of Radiation Oncology, CARTI Cancer Center, Little Rock, AR, United States of America

[3]Department of Gynecologic Oncology, Brown Cancer Center, University of Louisville School of Medicine, Louisville, KY, United States of America

[4]Department of Bioinformatics & Biostatistics, University of Louisville School of Public Health and Information Sciences, Louisville, KY, United States of America

[5]Department of Radiation Oncology, Vanderbilt University Medical Center, Nashville, TN, United States of America

* Corresponding author

E-mail: aportocarrerob@pucp.edu.pe

# Abstract

**Background**:

Gynecological cancers are among the most prevalent cancers in women worldwide. Brachytherapy, often used as a boost to external beam radiotherapy, is integral to treatment. Advances in computation, algorithms, and data availability have popularized machine learning.

**Objective**:

To develop and compare machine learning models for predicting grade 3 or higher toxicities in gynecological cancer patients treated with high dose rate (HDR) brachytherapy, aiming to contribute to personalized radiation treatments.

**Methods**:

A retrospective analysis on gynecological cancer patients who underwent HDR brachytherapy with Syed-Neblett or Tandem and Ovoid applicators from 2009 to 2023. After exclusions, 233 patients were included. Dosimetric variables for the high-risk clinical target volume (HR-CTV) and organs at risk, along with tumor, patient, and toxicity data, were collected and compared between groups with and without grade 3 or higher toxicities using statistical tests. Six supervised classification machine learning models (Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machines, Gaussian Naive Bayes, and Multi-Layer Perceptron Neural Networks) were constructed and evaluated. The construction process involved sequential feature selection (SFS)

when appropriate, followed by hyperparameter tuning. Final model performance was characterized using a 25% withheld test dataset.

**Results**:

The top three ranking models were Support Vector Machines, Random Forest, and Logistic Regression, with F1 testing scores of 0.63, 0.57, and 0.52; normMCC testing scores of 0.75, 0.77, and 0.71; and accuracy testing scores of 0.80, 0.85, and 0.81, respectively. The SFS algorithm selected 10 features for the highest-ranking model. In traditional statistical analysis, HR-CTV volume, Charlson Comorbidity Index, Length of Follow-Up, and D2cc - Rectum differed significantly between groups with and without grade 3 or higher toxicities.

**Conclusions**:

Machine learning models were developed to predict grade 3 or higher toxicities, achieving satisfactory performance. Machine learning presents a novel solution to creating multivariable models for personalized radiation therapy care.

# Introduction

Gynecological cancers rank among the most diagnosed malignancies affecting women on a global scale [1]. In the United States of America, it is estimated that there will be 116,930 new cases and 36,250 deaths in 2024 attributable to gynecologic malignancies [2]. Treatments for gynecologic cancers include surgery, chemotherapy, and/or radiotherapy [3]. Brachytherapy is necessary in the management of locally advanced cervical cancer, since patients who do not receive brachytherapy following concurrent

63 external beam radiation therapy and chemotherapy have significantly worse overall
64 survival [4]. Colson-Fearon et al. reported that the 4-year overall survival in locally
65 advanced cervical cancer patients treated with brachytherapy versus without
66 brachytherapy is 67.7% versus 45.7%, respectively [5]. With 3-dimensional magnetic
67 resonance image-guided brachytherapy for cervical cancer, the 5-year local control is
68 92% [6].
69
70 Machine learning (ML) has been defined as an optimization problem to find the most
71 suitable predictive model for new data based on an existing dataset obtained from a similar
72 context [7]. The recent rise in popularity of ML has been due to the development of new
73 algorithms, theory, data availability, and improvements in low-cost computation [8]. For
74 many problems, ML has shown to have better overall predictive metrics than conventional
75 statistical models (CSM) [9-11].
76
77 ML is a bottom-up approach that has the advantages of being data-driven, of not requiring
78 strict a-priori assumptions about the forms of the relationships between variables and
79 outcomes, and of accounting for complex interactions among input features. In contrast,
80 CSMs can be viewed as top-down approaches, and their main advantages are their
81 interpretability due to usually focusing on the parsimonious relationships between input
82 and response, the low computational resources needed to fit the models, and being less
83 susceptible to overfitting with large datasets [12-13].
84

Binary classification, in which the ML model predicts an output that is either one of two possible classes, is one of the most common tasks that can be solved with supervised machine learning [14]. For this problem, a model is trained with data that contains both features and the response labels, and the algorithm compares the actual and predicted results using an appropriate assessment metric [15]. This study aims to build and compare some of the more common binary classification machine learning models in the context of predicting if a patient is going to develop grade 3 or higher toxicities (Output: Yes/No) in gynecologic cancer patients treated with EBRT and brachytherapy.

# Methods

## Data Collection

A comprehensive retrospective analysis was conducted, encompassing a total of 233 patients who had undergone high dose rate (HDR) brachytherapy with Syed-Neblett or Fletcher-Suit-Delclos Tandem and Ovoid (T&O) applicators for treatment of gynecological cancer (cervix, endometrium, vagina, or vulva) at a single institution spanning the period from 2009 to 2023. Demographic details, tumor characteristics, treatment variables, dosimetric information (including if the patient received an EBRT boost), and occurrences of gastrointestinal (GI), genitourinary (GU), and vaginal (VAG) toxicities during and post-treatment were gathered. The exclusion criteria included the following: patients with a prior brachytherapy history, those treated with more than a single type of brachytherapy applicator, conflicting dosimetric data found in records, concurrent external beam radiotherapy for a distinct proximal disease site, or a combination of low dose rate (LDR)

and HDR treatments. Toxicities were classified according to the Common Terminology Criteria for Adverse Events (CTCAE) version 5.0 [16], and the integrity of the database was reviewed three times by both a physician and a medical physicist to ensure its accuracy and reliability. For treatment planning, the dosimetry goals as detailed in the EMBRACE trials and ASTRO Clinical Practice Guideline were followed [6,17]. All patients received EBRT and Brachytherapy. The process used to calculate the total EQD2 dose has been described in-detail in a previous work, and follows the procedure suggested by ICRU 89 [18-19]. This study was approved by our institutional review board (IRB 22.0117).

## Statistical Analysis

Preliminary dataset exploration was done by comparing between patients that developed no higher than a grade 2 toxicity and those that developed grade 3 or higher toxicities at any point in time after EBRT initiation; continuous variables were reported as means and standard deviations and compared with 2-sample t-tests. Categorical variables were listed as counts and percentages and compared with the Fisher exact test. Non-normal continuous variables were reported at median and interquartile range (IQR) and compared with the non-parametric Mann-Whitney test; a p-value of 0.05 or lower was considered to be statistically significant. Kaplan-Meier curves for disease free survival and local control were created to characterize the cohort.

## Data Preprocessing

The analysis was done using Python 3 and Jupyter Notebook (IPython kernel). Various code libraries (collections of pre-written functions and classes), including Scikit-learn v1.3.2 [20], were used for their efficiency and reproducibility; care was taken to ensure compatibility and the use of the appropriate library versions. Charlson Comorbidity Index was categorized into approximate quartiles ("Low" (0-2), "Medium" (3), "High" (4-5), or "Very High" ( > 5)), and KPS was assigned categories according to clinical interpretation: "Bad" (50-70), "Normal" (80), or "Good" (90-100). Data pre-processing involved four steps: A) Encoding, B) Imputation, C) Class Balancing and D) Normalization.

For data encoding, categorical and ordinal variables were assigned to numeric labels. The data then underwent a stratified split based on the target, resulting in two groups with an equivalent proportion of toxicity events: 75% for training (n = 174) and 25% for testing (n = 59).

Imputation of missing feature values was done according to the variable type. For categorical and ordinal variables, a K-nearest neighbors (KNN) imputer was employed utilizing the single nearest neighbor to guarantee imputation to a single class for that feature. For the numerical continuous features, the KNN imputer was used with K = 5 neighbors, and the missing features were imputed by the average. This parameter was chosen after extensive experimentation. These imputers identify their nearest neighbors by calculating the Euclidean Distance between data points (not including the missing data). They were fitted using the training data only, and their algorithms applied to both the training and testing data [21].

The defined positive class of Grade 3 or higher toxicity was observed in a minority of patients (24%, 56/233), leading to an imbalanced dataset. To address this imbalance, the class-balancing algorithm SVM-SMOTE [22] was used only during model training. (Preliminary analyses suggested this algorithm had better performance than alternative balancing algorithms such as SMOTE [23] and ADASYN [24]). Out of the initial 174 samples from the training data, an additional 90 synthetic positive cases were generated for a total of 264 samples (132 positive, 132 negative).

The final pre-processing step included the normalization/standardization of values. After experimentation the Standard Scaler was used for continuous numerical variables. For categorical and ordinal variables, Target Encoding was used. Other normalization/standardization methods such as MinMax Scaler and the Robust Scaler were also explored but not reported in this work due to obtaining worse results. The fitted Scalers and the Target Encoding objects were stored into a Joblib file and then employed in the testing data.

Investigation into collinearity between input features was also performed using Pearson's correlation coefficient. The final model eliminated one of the pairs of collinear features with values greater than 0.80 correlation. Other thresholds such as 0.7 and 0.95 were also analyzed but yielded worse results. When dose metrics were collinear, D2cc and D90 were given priority to remain in the final model due to being the most widely used clinical values [17].

## Evaluation of Machine Learning Models

There are multiple performance metrics (PM) that can be used to assess a model performance on predicting new data. In this study, the Accuracy, Precision, Recall, F1 score, Matthews Correlation Coefficient (MCC), the area under the curve of a receiver operating characteristic curve (AUC-ROC) and the area under the curve of a precision-recall curve (AUC-PR) were used; the first four metrics are defined using the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1) \qquad Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3) \qquad F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (5)$$

In this context, Positive/Negative represents whether the ML model predicts a toxicity event, and True/False represents whether the ML prediction agrees/disagrees with the patient record. Accuracy as shown in formula (1) is the ratio of correctly predicted instances over the total number of patients. Precision, which is represented by formula (2), is the ratio of correctly predicted positive observations to the total number of observations predicted to be positive. Recall, also known as Sensitivity, is the ratio of correct predictions among patients with toxicities as shown in formula (3); the F1 score, as defined in formula (4), is equivalent to the harmonic mean of precision and recall [25]; MCC, or its normalized version (normMCC) [26], is a balanced measure that considers all four basic metrics (TP, FP, TN, FN) as shown in formula (5). Additional metrics such as the AUC-ROC and AUC-PR evaluate the overall performance of the model by

considering performance across all possible decision thresholds of the model [27]. In this work, the reported F1, recall, and precision scores are calculated for the positive class (patients that present a toxicity). For the AUC-ROC curve, the baseline denotes a random classifier, manifesting as a diagonal line with an AUC-ROC value of 0.5. Conversely, the PR curve's baseline reflects a situation where all classifications are assumed to be positive, resulting in a horizontal line on the precision-recall plot; the position of this line on the Y-axis is contingent upon the characteristics of the data under consideration [28-29]. These prediction metrics are calculated and reported for the training (without the SVM-SMOTE generated synthetic samples used for data balancing) and withheld test data (with the missing data-imputed for both). For the purposes of this work the authors have considered the top ML models as the ones with the highest test data F1 score. Confidence intervals of 95% were calculated assuming a normal distribution, as justified by the Central Limit Theorem [30].

$$PM \pm 1.96\sqrt{\frac{1}{n} \cdot PM \cdot (1 - PM)}$$

## Sequential Feature Selection

In various domains, including healthcare, datasets may exhibit high dimensionality, referring to the presence of a large number of variables or features. This characteristic can adversely affect the development and interpretability of some machine learning algorithms (Logistic Regression, Support Vector Machines, K Nearest Neighbors, and Gaussian Naive Bayes) [31-32]. To reduce dimensionality, several approaches exist such as feature extraction and feature selection [33]. In this work, multiple variations of sequential feature selection were initially considered, including Sequential Forward

Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS) and Sequential Backward Floating Selection (SBFS), which used as their estimator the same model to later be used for training [34-36]; after experimentation, Sequential Forward Selection was chosen for the full analysis due to faster computation time and better performance metrics. Note that, regardless of traditional statistical analysis, both marginally significant variables, and those that were not, are explored when training the ML algorithms. The forward feature selection process adds one feature into the model at a time, determining inclusion based on which predictor optimizes the evaluation criteria, which in our case was the F1 score. As part of model training, a 10-fold Stratified Shuffle Split cross-validator was used over the class-balanced training data to reduce overfitting and appropriately assess the performance metrics of the sequential feature selection algorithm [37-38].

## Machine Learning Algorithms

A total of 6 machine learning models were implemented and compared. The included models were the following: Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Gaussian Naïve Bayes (GNB), and Multi-Layer Perceptron Neural Network (MLP). While there are many other ML classification algorithms in the literature, these six choices represent the most commonly utilized algorithms in this context. The baseline for the precision – recall curve was determined to be a horizontal line equal to 0.237 based on a classifier that labels all predictive instances as positive within the held-out testing data. After selecting the most relevant features through the Sequential Feature Selection process for the appropriate

models, the hyperparameters of all 6 models were further fine-tuned by using a Grid Search over another 10-fold Stratified Shuffle Split cross validator to optimize prediction under each model choice; the hyperparameter search space used by Grid Search is detailed in S1 Table. The Python code and a demonstration dataset of 50 randomly chosen patients have been made available to the reader. To safeguard patient privacy, utmost care was taken to avoid disclosing any identifiable health information. Moreover, noise was added to the demonstration dataset as an additional layer of protection. This data is available at: https://github.com/AndresPB95/ML-Model-Gynecological-HDR-G3Plus-Toxicities. A comprehensive diagram depicting the full machine learning workflow is provided in Fig 1, and S2 Table presents a summary of all the features explored by the ML models, along with their types.

**Fig 1**. **Flowchart outlining the steps used when training and evaluating the different models**. The process is divided in the following steps: **(A) Initial Train/Test split**: The data is initially divided into training and testing sets. The training set is used for most of the model development process, while the testing set is reserved to simulate new, unseen data. **(B) Data preprocessing (Training Set)**: Preprocessing steps include: (i) A *KNN Imputer* is fitted and applied to the training data to fill in missing values, (ii) *Collinear* features are removed, (iii) *SVM SMOTE* is used to oversample the positive class (*Only for training). Note: A separate, unbalanced copy of the training set was retained for evaluation, (iv) a *StandardScaler* is fitted and applied to the training data ensuring they are on a comparable scale. **(C) Data preprocessing (Testing Set)**: The preprocessing objects fitted to the training set are subsequently applied to the testing set: (i) The *KNN Imputer* is used to fill in missing values in the testing data (ii) (ii) *Collinear*

features are removed, (iii) the *StandardScaler* is applied for normalization. Note: SVM SMOTE was NOT used to oversample the test set. **(D) Hyperparameter Tuning**: For each model, the following tuning procedures are conducted using 10-fold cross-validation: (i) *Sequential Feature Selection* (if applicable) creates and trains multiple models by adding one feature at a time. Each model's F1 score is tested by comparing the predicted values with the known labels, and features that improve the F1 score are retained, building towards the most effective feature set. (ii) *GridSearch* trains multiple models with various hyperparameter combinations. Each combination's F1 score is tested by comparing the predicted values with the known labels, and the best-performing combination is selected for the final model. **(E) Final Model Generation**: After identifying the optimal hyperparameters and features, a final model is trained using the entire balanced training set. **(F) Evaluation**: The model's performance is evaluated by comparing its predictions against the known labels using both the unbalanced training set and the testing set.

# Results

The data included demographic and clinical data for n = 233 patients, of which n = 56 (24%) had a grade 3 or higher toxicity. The demographic, treatment, and tumor-related data are shown in Table 1. Patients who experienced grade 3 or higher toxicity were found to have longer follow-up (median 12.4 months versus 3.8 months), more likely to have low or very high comorbidity scores and had significantly higher HR-CTV values (median 50 cc versus 39 cc, p = 0.041).

**Table 1**. **Comparison of patient, treatment, and tumor characteristics between groups with and without grade 3 or higher toxicities**.

| | Full Cohort | | No Grade 3+ Toxicity | | Grade 3+ Toxicity | | |
|---|---|---|---|---|---|---|---|
| | n=233 | 100% | n=177 | 76% | n=56 | 24% | p-value |
| Length of Follow-Up (mo) | 6.1 | IQR: [1.4 - 18.2] | 3.8 | IQR: [1.2 - 16.4] | 12.4 | IQR: [7.1 - 22.1] | ***< 0.001*** |
| Age at Completion | 53.6 | STD: 14.8 | 54.4 | STD: 14.7 | 50.8 | STD: 14.8 | 0.107 |
| Non-Caucasian | 25 | 11% | 17 | 10% | 8 | 14% | 0.328 |
| BMI | 28.0 | STD: 8.3 | 28.0 | STD: 8.6 | 27.9 | STD: 7.6 | 0.969 |
| Charlson Comorbidity Index | | | | | | | ***0.014*** |
| Low [0-2] | 87 | 37% | 60 | 34% | 27 | 48% | |
| Medium [3] | 43 | 18% | 33 | 19% | 10 | 18% | |
| High [4-5] | 60 | 26% | 54 | 31% | 6 | 11% | |
| Very High [>5] | 43 | 18% | 30 | 17% | 13 | 23% | |
| KPS | | | | | | | 0.369 |
| Good [90-100] | 147 | 63% | 115 | 65% | 32 | 57% | |
| Normal [80] | 62 | 27% | 43 | 24% | 19 | 34% | |
| Bad [50-70] | 23 | 10% | 18 | 10% | 5 | 9% | |
| Treatment Days | 60 | IQR: [52 - 71] | 60 | IQR: [52 - 69] | 61 | IQR: [52 - 74] | 0.504 |
| Applicator: T&O | 80 | 34% | 63 | 36% | 17 | 30% | 0.521 |
| Concurrent Chemo | 201 | 86% | 153 | 86% | 48 | 86% | 1.000 |
| Type of Boost | | | | | | | 0.681 |
| None | 139 | 60% | 108 | 61% | 31 | 55% | |
| Sequential | 54 | 23% | 39 | 22% | 15 | 27% | |
| SIB | 40 | 17% | 30 | 17% | 10 | 18% | |
| Tumor Size (cm) | 5.4 | STD: 2.1 | 5.4 | STD: 2.0 | 5.6 | STD: 2.5 | 0.622 |
| HR-CTV (cc) | 43 | IQR: [27 - 74] | 39 | IQR: [25 - 71] | 50 | IQR: [34 - 77] | ***0.041*** |
| Tumor Site | | | | | | | 0.864 |
| Cervix | 194 | 83% | 147 | 83% | 47 | 84% | |
| Endometrium | 16 | 7% | 13 | 7% | 3 | 5% | |
| Other | 23 | 10% | 17 | 10% | 6 | 11% | |
| Cancer Stage | | | | | | | 0.163 |
| Stage 1 | 45 | 19% | 37 | 21% | 8 | 14% | |
| Stage 2 | 57 | 25% | 41 | 23% | 16 | 29% | |
| Stage 3 | 107 | 46% | 84 | 48% | 23 | 41% | |
| Stage 4 | 22 | 10% | 13 | 7% | 9 | 16% | |
| Histology: SCC | 180 | 77% | 136 | 77% | 44 | 79% | 0.857 |
| MRI Fused | 105 | 45% | 84 | 47% | 21 | 38% | 0.219 |

Table 2 compares median dose coverage to the tumor (V100%, D50%, D90%, and D98%) and the dose to the organs at risk (OARs) by toxicity status. Patients with toxicities

had significantly higher D2cc doses to the rectum (p = 0.043), but no other doses were statistically significantly different. The HR-CTV V100, D1cc - Rectum, and doses to the sigmoid colon were slightly higher for the group with grade 3 or higher toxicities but not statistically significant.

**Table 2. HR-CTV and OAR dosimetric values between groups with and without grade 3 or higher toxicities.**

| | Full Cohort N=233 (100%) | | No Grade 3+ Toxicity N=177 (76%) | | Grade 3+ Toxicity N=56 (23%) | | |
|---|---|---|---|---|---|---|---|
| | Median | IQR | Median | IQR | Median | IQR | p-value |
| HR-CTV V100 (cc) | 80.1 | [54.3 - 129.3] | 74.5 | [50.8 - 123.3] | 88.3 | [60.3 - 132.7] | 0.104 |
| HR-CTV D50 (Gy) | 110.0 | [101.2 - 119.1] | 110.8 | [101.4 - 119.1] | 109.8 | [100.3 - 118.9] | 0.628 |
| HR-CTV D90 (Gy) | 83.1 | [79.9 - 87.7] | 83.1 | [80.0 - 87.7] | 83.1 | [79.2 - 86.7] | 0.967 |
| HR-CTV D98 (Gy) | 75.3 | [70.9 - 79.9] | 75.1 | [70.9 - 79.8] | 75.7 | [69.7 – 80.0] | 0.837 |
| D0.1cc - Bladder (Gy) | 97.5 | [83.4 - 114.5] | 97.0 | [83.3 - 112.9] | 98.6 | [85.2 - 120.8] | 0.455 |
| D1cc - Bladder (Gy) | 84.9 | [74.9 - 95.4] | 84.6 | [75.2 - 94.7] | 85.5 | [74.6 - 98.1] | 0.570 |
| D2cc - Bladder (Gy) | 79.7 | [71.3 - 89.2] | 79.2 | [71.8 - 88.5] | 81.9 | [70.5 - 91.1] | 0.569 |
| D0.1cc - Small Bowel (Gy) | 67.2 | [52.8 - 83.8] | 68.0 | [52.7 - 84.5] | 65.2 | [53.9 - 78.2] | 0.838 |
| D1cc - Small Bowel (Gy) | 60.9 | [51.5 - 73.5] | 61.1 | [51.2 - 73.7] | 59.6 | [52.4 - 70.1] | 0.852 |
| D2cc - Small Bowel (Gy) | 59.3 | [50.7 - 69.6] | 59.4 | [50.6 – 70.0] | 57.8 | [51.0 - 66.4] | 0.898 |
| D0.1cc - Sigmoid Colon (Gy) | 74.3 | [62.0 – 86.0] | 73.9 | [59.9 - 86.3] | 76.6 | [65.2 - 85.4] | 0.277 |
| D1cc - Sigmoid Colon (Gy) | 67.0 | [57.2 - 75.4] | 66.5 | [56.1 – 75.0] | 69.1 | [60.3 - 76.3] | 0.182 |
| D2cc - Sigmoid Colon (Gy) | 64.2 | [55.3 - 71.6] | 63.8 | [54.3 - 71.2] | 65.6 | [58.0 - 72.2] | 0.172 |
| D0.1cc - Rectum (Gy) | 80.8 | [74.3 - 87.6] | 80.7 | [74.0 - 87.6] | 82.3 | [76.8 - 86.9] | 0.361 |
| D1cc - Rectum (Gy) | 71.4 | [66.2 - 78.4] | 71.1 | [65.7 - 76.6] | 73.7 | [68.5 - 79.9] | 0.066 |
| D2cc - Rectum (Gy) | 67.6 | [62.5 - 74.4] | 67.3 | [62.2 - 72.6] | 71.1 | [65.3 - 75.9] | ***0.043*** |

The six machine learning models were then fitted using all variables included in Table 1 and Table 2 as described in the Methods section. The performance of these models on the withheld test data are depicted visually in Figs 2 and 3. Numeric comparisons based on both the (class-imbalanced) training data and withheld test data are shown in Table 3. The top three models for predicting grade 3 or higher toxicities are found to be Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR) with F1 testing scores of 0.63, 0.57 and 0.52, normMCC testing scores of 0.75, 0.77 and 0.71, and Accuracy testing scores of 0.80, 0.85 and 0.81, respectively. All values shown in Table 3 assume a classification threshold value of 0.5 for toxicity prediction. Note that this table also includes the metrics from the training data, which for some models (MLP and KNN) disagree strongly with the test data performance measures, indicating severe overfitting in the training data. Table 4 exhibits the most relevant features and the values of the hyperparameters selected by the GridSearchCV optimization algorithm over the training data. The top features repeated among these three models are Chemotherapy, Charlson Comorbidity Index, KPS, D2cc - Small Bowel, Stage, Histology, and Follow-Up Time.

**Fig 2**. **Precision-Recall curves comparing 6 machine learning models and a baseline value**. PR curves are computed using the withheld test data. SVM is the model with the highest area under the curve.

**Fig 3. Receiver Operating Characteristics curves for 6 machine learning models and a baseline value**. ROC curves are computed using the withheld test data. SVM is the model with the highest area under the curve.

**Table 3. Training and testing performance metrics for the considered machine learning models**.

| Dataset | Model | F1 | | Accuracy | | normMCC | | Precision | | Recall | | AUC-ROC | | AUC-PR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | 95% CI | Mean | 95% CI | Mean | 95% CI | Mean | 95% CI | Mean | 95% CI | Mean | 95% CI | Mean | 95% CI |
| Training | SVM | 0.61 | 0.54 - 0.68 | 0.82 | 0.76 - 0.87 | 0.75 | 0.68 - 0.81 | 0.63 | 0.55 - 0.70 | 0.60 | 0.52 - 0.67 | 0.87 | 0.82 - 0.92 | 0.60 | 0.52 - 0.67 |
| | RF | 0.82 | 0.76 - 0.88 | 0.92 | 0.88 - 0.96 | 0.89 | 0.84 - 0.93 | 0.89 | 0.84 - 0.94 | 0.76 | 0.70 - 0.83 | 0.97 | 0.95 - 1.00 | 0.93 | 0.89 - 0.96 |
| | LR | 0.45 | 0.38 - 0.52 | 0.75 | 0.68 - 0.81 | 0.64 | 0.57 - 0.71 | 0.47 | 0.40 - 0.55 | 0.43 | 0.36 - 0.50 | 0.72 | 0.65 - 0.79 | 0.43 | 0.36 - 0.51 |
| | MLP | 1.00 | 1.00 - 1.00 | 1.00 | 1.00 - 1.00 | 1.00 | 1.00 - 1.00 | 1.00 | 1.00 - 1.00 | 1.00 | 1.00 - 1.00 | 1.00 | 1.00 - 1.00 | 1.00 | 1.00 - 1.00 |
| | KNN | 0.72 | 0.65 - 0.79 | 0.86 | 0.8 - 0.91 | 0.81 | 0.75 - 0.87 | 0.68 | 0.61 - 0.75 | 0.76 | 0.70 - 0.83 | 0.91 | 0.87 - 0.95 | 0.78 | 0.72 - 0.84 |
| | GNB | 0.33 | 0.26 - 0.40 | 0.79 | 0.73 - 0.85 | 0.66 | 0.59 - 0.73 | 0.75 | 0.69 - 0.81 | 0.21 | 0.15 - 0.28 | 0.67 | 0.60 - 0.74 | 0.40 | 0.33 - 0.47 |
| Testing | SVM | 0.63 | 0.50 - 0.75 | 0.80 | 0.69 - 0.90 | 0.75 | 0.64 - 0.89 | 0.56 | 0.43 - 0.68 | 0.71 | 0.60 - 0.83 | 0.78 | 0.67 - 0.88 | 0.65 | 0.53 - 0.77 |
| | RF | 0.57 | 0.45 - 0.70 | 0.85 | 0.76 - 0.94 | 0.77 | 0.66 - 0.88 | 0.86 | 0.77 - 0.95 | 0.43 | 0.30 - 0.55 | 0.76 | 0.65 - 0.87 | 0.52 | 0.39 - 0.65 |
| | LR | 0.52 | 0.39 - 0.65 | 0.81 | 0.71 - 0.91 | 0.71 | 0.60 - 0.83 | 0.67 | 0.55 - 0.79 | 0.43 | 0.30 - 0.55 | 0.68 | 0.56 - 0.80 | 0.47 | 0.34 - 0.60 |
| | MLP | 0.39 | 0.26 - 0.51 | 0.63 | 0.50 - 0.75 | 0.57 | 0.45 - 0.70 | 0.32 | 0.20 - 0.44 | 0.50 | 0.37 - 0.63 | 0.66 | 0.54 - 0.78 | 0.44 | 0.31 - 0.56 |
| | KNN | 0.32 | 0.20 - 0.44 | 0.64 | 0.52 - 0.77 | 0.54 | 0.42 - 0.67 | 0.29 | 0.18 - 0.41 | 0.36 | 0.23 - 0.48 | 0.62 | 0.50 - 0.74 | 0.46 | 0.33 - 0.58 |
| | GNB | 0.24 | 0.13 - 0.34 | 0.78 | 0.67 - 0.89 | 0.62 | 0.49 - 0.74 | 0.67 | 0.55 - 0.79 | 0.14 | 0.05 - 0.23 | 0.57 | 0.44 - 0.69 | 0.34 | 0.22 - 0.46 |

**Table 4. Most important features as selected by the Sequential Feature Selection algorithm (where appropriate) and found optimal hyperparameters for the top 3 scoring models**.

| SVM | | RF | | LR | |
|---|---|---|---|---|---|
| Features | Hyperparameters | Features | Hyperparameters | Features | Hyperparameters |
| Chemotherapy | C: 1 | All | n_estimators: 15 | Chemotherapy | C: 1 |
| Charlson | kernel: rbf | | max_features: log2 | Charlson | penalty: 50 |
| KPS | gamma: scale | | min_samples_leaf: 5 | KPS | solver: lbfgs |
| MRI | | | min_samples_split: 5 | Ethnicity | |
| D2cc Small Bowel | | | | Type of Boost | |
| D2cc Sigmoid | | | | Applicator | |
| Stage | | | | D2cc Small Bowel | |
| Histology | | | | D2cc Rectum | |
| HR-CTV | | | | Tumor Site | |
| Follow-Up Time | | | | Stage | |
| | | | | Histology | |
| | | | | Follow-Up Time | |

# Discussion

This study aimed to investigate the utility of using machine learning models to predict grade 3 or higher toxicities in gynecologic cancer patients treated with EBRT and interstitial or T&O brachytherapy. The database was analyzed using traditional statistics which compared groups with and without grade 3+ toxicities; disease free survival and local control were also reported (S1 Fig). To design the toxicity prediction models, data

were encoded and pre-processed. Next, a sequential feature selector method was used when appropriate, and hyperparameter tuning was performed.

A comparison of the patients with and without grade 3 toxicities, using basic marginal statistical analysis, suggested few differences between the groups including HR-CTV, Charlson Comorbidity Index, Length of Follow Up, and D2cc - Rectum. Some of these variables such as the HR-CTV and D2cc - Rectum have been previously shown to be predictors of grade 3 or higher toxicities for HDR brachytherapy. Lee et al. observed that patients with grade 3-4 toxicities had a significantly higher median HR-CTV of 111 cc compared to 43 cc for those patients with grade 0-2 toxicities [39]. Mesko et al. found a statistically significant difference between patients with and without a grade 3 toxicity, with a median of 93.8 cc and 51 cc, respectively [40]. Mazeron et al. found that rectal D2cc values equal to or greater than 75 Gy EQD2 are associated with higher grade and more frequent toxicities in MRI-guided adaptive brachytherapy for locally advanced cervical cancer [41]. When compared with traditional statistics, machine learning models consider nonlinear interactions between variables [42], resulting in our top scoring model selecting a total of 10 features. One should keep in mind that the practical importance of each feature within an ML algorithm may vary and their relevance to the outcome should not be inferred solely based on their inclusion in the model. Additionally, the features chosen by SFS may exclude variables that are easily manipulable when creating a treatment plan, particularly dosimetric variables. This issue could be explained twofold: 1) certain combinations of hyperparameters could limit the ability of SFS to find the correct interactions between features in the final selection; or 2) certain combinations of features could be more relevant and produce better predictions than when using actual dosimetric

data. A model without any dosimetric features would still be useful for predicting toxicity risk, but would not provide the clinician the option of adjusting the treatment plan to reduce the risk of toxicity.

Supervised machine learning has been utilized to perform classification tasks in various areas of healthcare such as for predicting diagnosis and prognosis of COVID-19 patients, prediction of hospitalization due to heart disease, and outcome prediction of infectious diseases [43-45]. To the authors' knowledge, this is the first analysis using and comparing multiple models for predicting grade 3 or higher toxicities in gynecologic cancer patients treated with external beam radiation and HDR interstitial or T&O brachytherapy. Through March 2020, there were only 53 published studies on the use of machine learning to predict radiation-induced toxicities [46], and through September 2023, only 14 studies had been published on deep learning models to predict toxicities from radiation treatment [47].

Regarding ML in brachytherapy toxicity prediction, Tian et al. developed a model for predicting fistula formation, reporting a recall of 97.1% and AUC of 0.904 utilizing the SMOTE algorithm and a SVM model with a radial basis kernel function on a database that included 35 patients with 7 positive cases; the limitation of this study lies in the small dataset, no withheld test dataset, high risk of model overfitting, and only using one model in their study [48]. For prediction of rectal toxicities, Chen et al. and Zhen et al. predicted grade 2 or higher rectal toxicity by using SVM and convolutional neural networks, respectively, with scores of (cross-validation estimated) recall and AUC of 0.85 and 0.91 for the former and 0.75 and 0.89 for the latter. Their work includes the addition of dose map features for the training of the model; both of these works were done with a database

of 42 patients with 12 positive cases of patients that developed toxicities [49-50]. Additionally, there has been work by Lucia et al. who developed Normal Tissue Complication Probability (NTCP) models for acute and late gastrointestinal, genitourinary, and vaginal toxicities using a database of 102 patients that included radiomic features, but only for a logistic regression model, which obtained balanced accuracy scores between 63.99 and 78.41 [51]. Cheon et al. considered deep learning models for predicting late bladder toxicities which outperformed its multivariable logistic regression counterpart [52], with data of 281 patients which achieved an F1 score of 0.76. In contrast to the preceding studies, our study presents the largest patient dataset used for predicting grade 3 or higher toxicities. Similar to these studies, we employ data-balancing algorithms to promote stability in the model training stage. Our methodology incorporates feature selection for all models except for MLP and RF. Specifically, we leverage the Sequential Feature Selection Algorithm to promote parsimony within the model fit. This aligns with the methodologies employed in previous reports.

Overfitting occurs when a model becomes overly complex, capturing noise in the training data instead of learning the underlying patterns, leading to poor predictions when applied to new data [53]. To mitigate this phenomenon, the use of a withheld testing data set is required to assess the degree of overfitting and the performance of the model [54]. A clear illustration of overfitting can be appreciated in Table 3 for the MLP and KNN models, where they achieved impressive training F1 scores of 1.00 and 0.72 respectively; contrasting sharply with their testing scores of 0.39 and 0.32. These scores show that these 2 models are not generalizable for predicting new similar data points. Further model exploration with an expansion of the hyperparameter search space and pre-processing

algorithms is needed and will be taken into account in future projects. The authors suggest that the training and withheld data testing scores are always reported for a comprehensive understanding of a model's performance.

Regarding the scoring metrics, our study showed that the support vector machine was the best model for predicting grade 3 toxicities, obtaining a training F1 score of 0.61, accuracy of 0.82, normMCC of 0.75, precision of 0.63, recall of 0.6, AUC-ROC of 0.87, and AUC-PR of 0.6; whereas for that same model, the test data obtained an F1 score of 0.63, accuracy of 0.80, normMCC of 0.75, precision of 0.56, recall of 0.71, AUC-ROC of 0.78, and AUC-PR of 0.65. In the withheld test data, out of all the patients that had a toxicity (n = 14), 71% were correctly predicted by the model (TP = 10); and out of all the predicted cases, 56% represented a true toxicity event and were not false positives (FP = 8). Given the high level of uncertainty in whether patients will develop toxicities, this may be viewed as an adequate performance. An important detail that must be considered is that the precision value is as important as the recall, since during normal clinical practice it is equally as important to avoid false positives as it is to detect true positive cases. In particular, a toxicity prediction model may suggest that the physician consider lowering the dose to certain OARs to prevent these high-grade radiation-related side effects; an algorithm with good recall but prone to predicting false positives may lead to reducing the dose for a patient not susceptible to developing toxicities. This reduction, in turn, may involve sacrificing a portion of tumor coverage, potentially decreasing tumor control. For this reason, the F1 score emerges as the optimal metric for evaluating the model's performance. In future investigations within this area, prioritizing either the recall or the precision score, which is not replaceable by specificity, could be explored. Notably,

specificity becomes less valuable in situations marked by an imbalance with a majority of true negatives [55] as the model's ability to predict negative outcomes can render overly optimistic scores in such scenarios. Once a best performing model has been identified, multi-institutional clinical trials will be needed to assess their performance on routine clinical practice.

The strength of this work lies in several key aspects. First, the study analyzes multiple machine learning models to find the best fit across a variety of common prediction algorithms. Additionally, we divide the data into training and testing sets before employing cross-validation for the model’s training, enhancing generalizability of the final models and providing more trustworthy measures of out-of-sample performance, despite potential reductions in the values of these metrics. The use of a Stratified Shuffle Split approach guarantees that there will be a positive class on the testing set of the cross validation, ensuring meaningful performance in every split. Furthermore, the focus on the F1 score and reporting precision as the performance metrics is of practical relevance for assessing the clinical performance of the model, especially when predicting toxicities.

The limitations are that, as in any machine learning study, having a larger dataset would likely help achieve better predictive accuracy, obtain a more generalizable model, and prevent overfitting. Additionally, only the dosimetric, treatment, and tumor variables were considered in this study, but not any additional features such as dose maps with spatial information. Regarding data balancing through Synthetic Oversampling, alternative techniques like threshold tuning could be investigated. Furthermore, developing methods to address overfitting and exploring a greater hyperparameter search space could be beneficial. Finally, an in-detail analysis of the importance of each hyperparameter could

be done in the future, using packages such as the Optuna library [56]; and additional ensemble models such as XGBoost could be trained and assessed. The authors acknowledge this and plan to address it in future studies.

## Conclusion

Multiple machine learning models were trained and assessed to predict grade 3 or higher toxicity development in patients with gynecologic malignancies who received EBRT and interstitial or T&O brachytherapy treatment yielding satisfactory results for the top performing model. This novel approach of toxicity prediction holds the potential to set a new paradigm in standard clinical care and contribute towards personalized care in radiation therapy. New techniques to improve model training need to be explored, and overcoming machine learning limitations like small datasets requires collaborative efforts among peers. In the future, further investigations are needed to prospectively validate these models in other healthcare settings.

# References

1. Costa M, Lai C. Coordinated efforts to harmonise gynaecological cancer care. *Lancet Oncol*. 2022;23(8):971-972. doi:10.1016/S1470-2045(22)00386-2
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74(1):12-49. doi:10.3322/caac.21830
3. Kehoe S. Treatments for gynaecological cancers. *Best Pract Res Clin Obstet Gynaecol*. 2006;20(6):985-1000. doi:10.1016/j.bpobgyn.2006.06.006

4. Robin TP, Amini A, Schefter TE, et al. Disparities in standard of care treatment and associated survival decrement in patients with locally advanced cervical cancer. *Gynecol Oncol*. 2016;143:319-25. doi:10.1016/j.ygyno.2016.09.009
5. Colson-Fearon D, Han K, Roumeliotis MB, et al. Updated Trends in Cervical Cancer Brachytherapy Utilization and Disparities in the United States From 2004 to 2020. *Int J Radiat Oncol Biol Phys*. 2024;119:154-62. doi:10.1016/j.ijrobp.2023.11.036
6. Potter R, Tanderup K, Schmid MP, et al. MRI-guided adaptive brachytherapy in locally advanced cervical cancer (EMBRACE-I): a multicentre prospective cohort study. *Lancet Oncol*. 2021;22:538-47. doi:10.1016/S1470-2045(20)30753-1
7. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis*. 2018;66(1):149-153. doi:10.1093/cid/cix731
8. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. 2015;349(6245):255-260.doi:0.1126/science.aac4520
9. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Fail*. 2021;8(1):106-115. doi:10.1002/ehf2.13073
10. Singal AG, Mukherjee A, Elmunzer BJ, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am J Gastroenterol*. 2013;108(11):1723-1730. doi:10.1038/ajg.2013.332
11. Chan K, Lee TW, Sample PA, et al. Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Trans Biomed Eng*. 2002;49(9):963-974. doi:10.1109/TBME.2002.802012
12. Rajula HSR, Verlato G, Manchia M, et al. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*. 2020;56(9):455. doi:10.3390/medicina56090455

13. Ley C, Martin RK, Pareek A, et al. Machine learning and conventional statistics: making sense of the differences. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(3):753-757. doi:10.1007/s00167-022-06896-6

14. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. *Artif Intell Rev*. 2006;26:159-190. doi:10.1007/s10462-007-9052-3

15. Saravanan R, Sujatha P. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2018;945-949. doi:10.1109/ICCONS.2018.8663155.

16. National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE). Available at: https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm. Accessed April 16, 2024.

17. Chino J, Annunziata CM, Beriwal S, et al. Radiation therapy for cervical cancer: executive summary of an ASTRO clinical practice guideline. *Pract Radiat Oncol*. 2020;10(4):220-234. doi:10.1016/j.prro.2020.04.002

18. Portocarrero-Bonifaz A, Syed S, Kassel M, et al. Dosimetric and toxicity comparison between Syed-Neblett and Fletcher-Suit-Delclos Tandem and Ovoid applicators in high dose rate cervix cancer brachytherapy. *Brachytherapy*. 2024;23(4):397-406. doi:10.1016/j.brachy.2024.03.003

19. International Commission on Radiation Units and Measurements. ICRU report 89: Prescribing, recording, and reporting brachytherapy for cancer of the cervix. *J ICRU*. 2016;13(1-2). doi:10.1093/jicru/ndw028

20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

21. Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explor*. 2001;3(1):27–32. doi:10.1145/507533.507538

22. Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. *Int J Knowl Eng Soft Data Paradigms*. 2011; 3(1), 4-21. doi: 10.1504/IJKESDP.2011.039875

23. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953

24. He H, Bai Y, Garcia EA, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008; 1322-1328. doi:10.1109/IJCNN.2008.4633969.

25. Vujović Ž. Classification model evaluation metrics. *Int J Adv Comput Sci Appl*. 2021;12(6):599-606. doi:10.14569/IJACSA.2021.0120670

26. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min*. 2023;16(1):4. doi:10.1186/s13040-023-00322-4

27. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159. doi:10.1016/S0031-3203(96)00142-2

28. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*. 2013;4(2):627.

29. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432

30. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv [Preprint]. 2018 [cited 2024 Aug 11]. Available from: https://arxiv.org/abs/1811.12808

31. Guyon I, Elisseeff A. An introduction to variable and feature selection. *JMLR*. 2003;3: 1157-1182. doi:10.1162/153244303322753616

32. Zekić-Sušac M, Pfeifer S, Nataša Šarlija N. A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem. *BSRJ*. 2014;5(3):82-96. doi:10.2478/bsrj-2014-0021

33. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432

34. Li J, Wang S, Morstatter F, et al. Feature selection: A data perspective. *ACM Comput Surv*. 2017;50(6):1-45. doi:10.1145/3136625

35. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. *Pattern Recognit Lett*. 1994;15(11):1119-1125. doi:10.1016/0167-8655(94)90127-9

36. Molina LC, Belanche L, Nebot À. Feature selection algorithms: A survey and experimental evaluation. *2002 IEEE International Conference on Data Mining, 2002. Proceedings*. 2002; 306-313. doi:10.1109/ICDM.2002.1183917.

37. Prusty S, Patnaik S, Dash SK. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front Nanotechnol*. 2022;4:972421. doi:10.3389/fnano.2022.972421

38. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J Anal Test*. 2018;2(3):249-262. doi:10.1007/s41664-018-0068-2

39. Lee LJ, Damato AL, Viswanathan AN. Clinical outcomes of high-dose-rate interstitial gynecologic brachytherapy using real-time CT guidance. *Brachytherapy*. 2013;12(4):303-310. doi:10.1016/j.brachy.2012.11.002

40. Mesko S, Swamy U, Park SJ, et al. Early clinical outcomes of ultrasound-guided CT-planned high-dose-rate interstitial brachytherapy for primary locally advanced cervical cancer. *Brachytherapy*. 2015;14(5):626-632. doi:10.1016/j.brachy.2015.04.006

41. Mazeron R, Fokdal LU, Kirchheiner K, et al. Dose-volume effect relationships for late rectal morbidity in patients treated with chemoradiation and MRI-guided adaptive brachytherapy for locally advanced cervical cancer: Results from the prospective multicenter EMBRACE study. *Radiother Oncol*. 2016;120:412-419. doi:10.1016/j.radonc.2016.06.006

42. Li R, Shinde A, Liu A, et al. Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival. *JCO Clin Cancer Inform*. 2020;4:637-646. doi:10.1200/CCI.20.00002

43. Muhammad LJ, Algehyne EA, Usman SS, et al. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Comput Sci*. 2021;2:1-13. doi:10.1007/s42979-020-00394-7

44. Dai W, Brisimi TS, Adams WG, et al. Prediction of hospitalization due to heart diseases by supervised learning methods. *Int J Med Inform*. 2015;84(3):189-197. doi:10.1016/j.ijmedinf.2014.10.002

45. Noorbakhsh-Sabet N, Zand R, Zhang Y, et al. Artificial intelligence transforms the future of health care. *Am J Med*. 2019;132(7):795-801. doi:10.1016/j.amjmed.2019.01.017

46. Isaksson LJ, Pepa M, Zaffaroni M, et al. Machine learning-based models for prediction of toxicity outcomes in radiotherapy. *Front Oncol*. 2020;10:790. doi:10.3389/fonc.2020.00790

47. Tan D, Nasir NFM, Manan HA, et al. Prediction of toxicity outcomes following radiotherapy using deep learning-based models: A systematic review. *Cancer Radiother*. 2023;27(5):398-406. doi:10.1016/j.canrad.2023.05.001

48. Tian Z, Zhou Z, Shen C, et al. A machine-learning–based prediction model of fistula formation after interstitial brachytherapy for locally advanced gynecological malignancies. *Brachytherapy*. 2019;18:530–8. doi:10.1016/j.brachy.2019.04.004

49. Chen J, Chen H, Zhong Z, et al. Investigating rectal toxicity associated dosimetric features with deformable accumulated rectal surface dose maps for cervical cancer radiotherapy. *Radiat Oncol*. 2018;13:125. doi:10.1186/s13014-018-1068-0

50. Zhen X, Chen J, Zhong Z, et al. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Phys Med Biol*. 2017;62:8246–63. doi:10.1088/1361-6560/aa8d09

51. Lucia F, Bourbonne V, Visvikis D, et al. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for cervical cancer. *J Pers Med*. 2021;11(5):398. doi:10.3390/jpm11050398

52. Cheon W, Han M, Jeong S, et al. Feature Importance Analysis of a Deep Learning Model for Predicting Late Bladder Toxicity Occurrence in Uterine Cervical Cancer Patients. *Cancers*. 2023;15(13):3463. doi:10.3390/cancers15133463

53. Peng Y, Nagata MH. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos Soliton Fract*. 2020; 139. doi:10.1016/j.chaos.2020.110055

54. El Naqa I, Boone JM, Benedict SH, et al. AI in medical physics: guidelines for publication. *Med Phys*. 2021; 48(9):4711-4714. doi:10.1002/mp.15170

55. Ali MM, Pau BK, Ahmed K, et al. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Comput Biol Med*. 2021;136:104672. doi:10.1016/j.compbiomed.2021.104672

56. Akiba T, Sano S, Yanase T, et al. Optuna: A next-generation hyperparameter optimization framework. KDD 19: 25th ACM SIGKDD international Conference on Knowledge Discovery & Data Mining. 2019; 2623-2631. doi:10.1145/3292500.3330701

# Supporting Information

**S1 Fig**. **Kaplan Meier plots for the entire patient cohort.** A) Disease Free Survival and B) Local control.

**S1 Table**. **Hyperparameter Search Space and MLP architecture**.

**S2 Table**. **Summary of input features and output from models**. The variable type and number of missing data points for each input is shown.

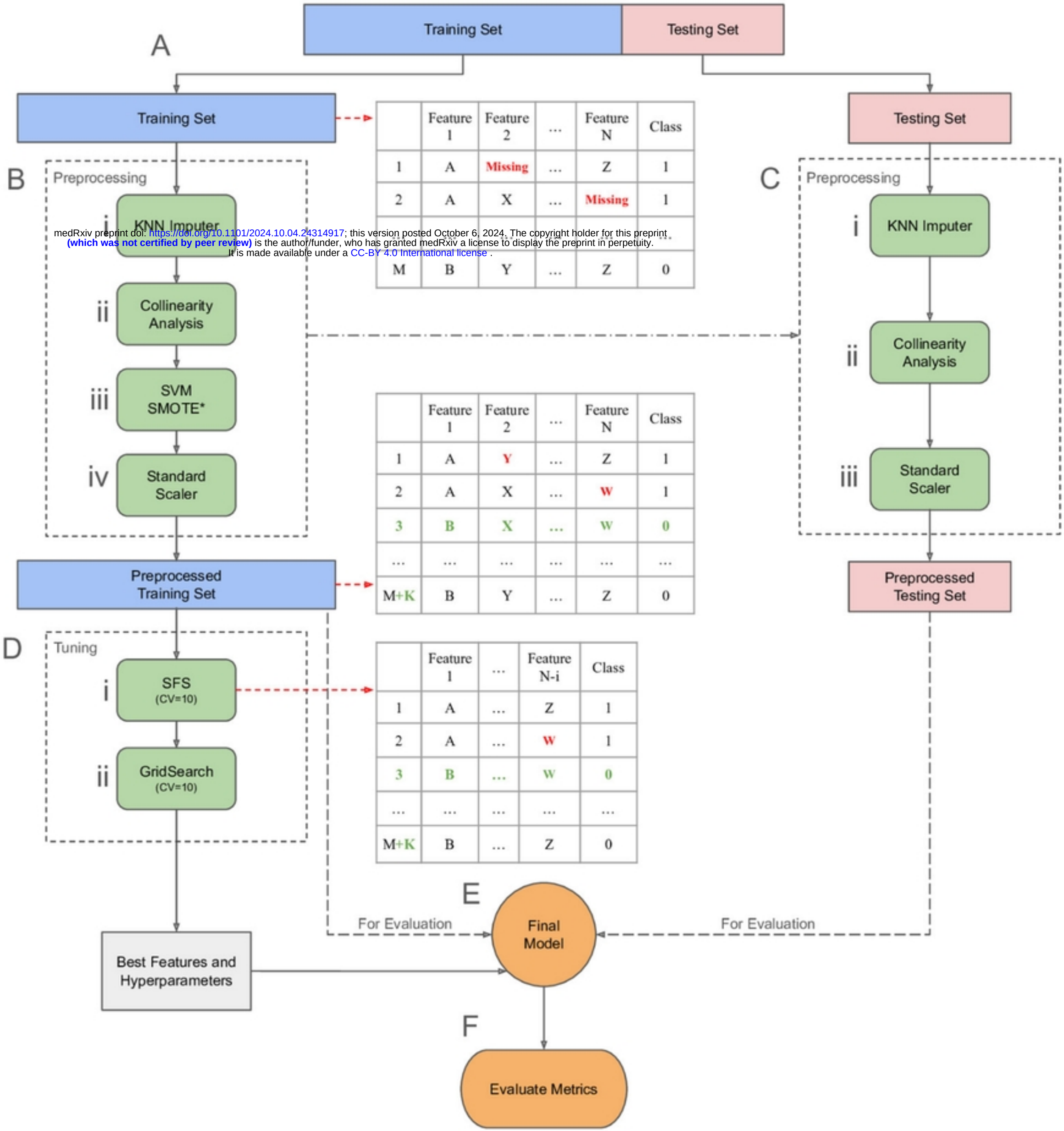


Figure 1

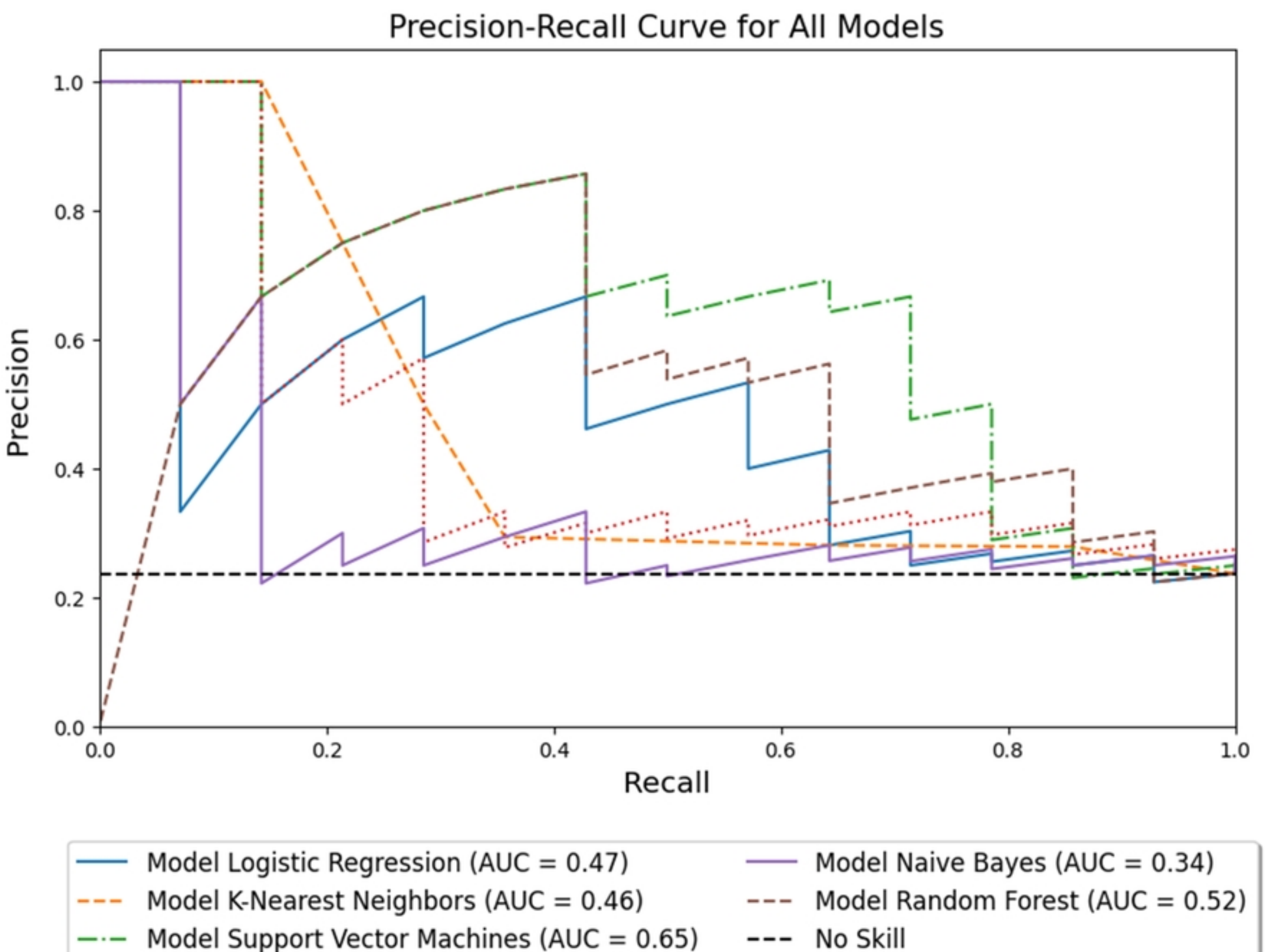
Precision-Recall Curve for All Models
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Model Logistic Regression (AUC = 0.47)
Model K-Nearest Neighbors (AUC = 0.46)
Model Support Vector Machines (AUC = 0.65)
Model MLP (AUC = 0.44)
Model Naive Bayes (AUC = 0.34)
Model Random Forest (AUC = 0.52)
No Skill

Figure 2

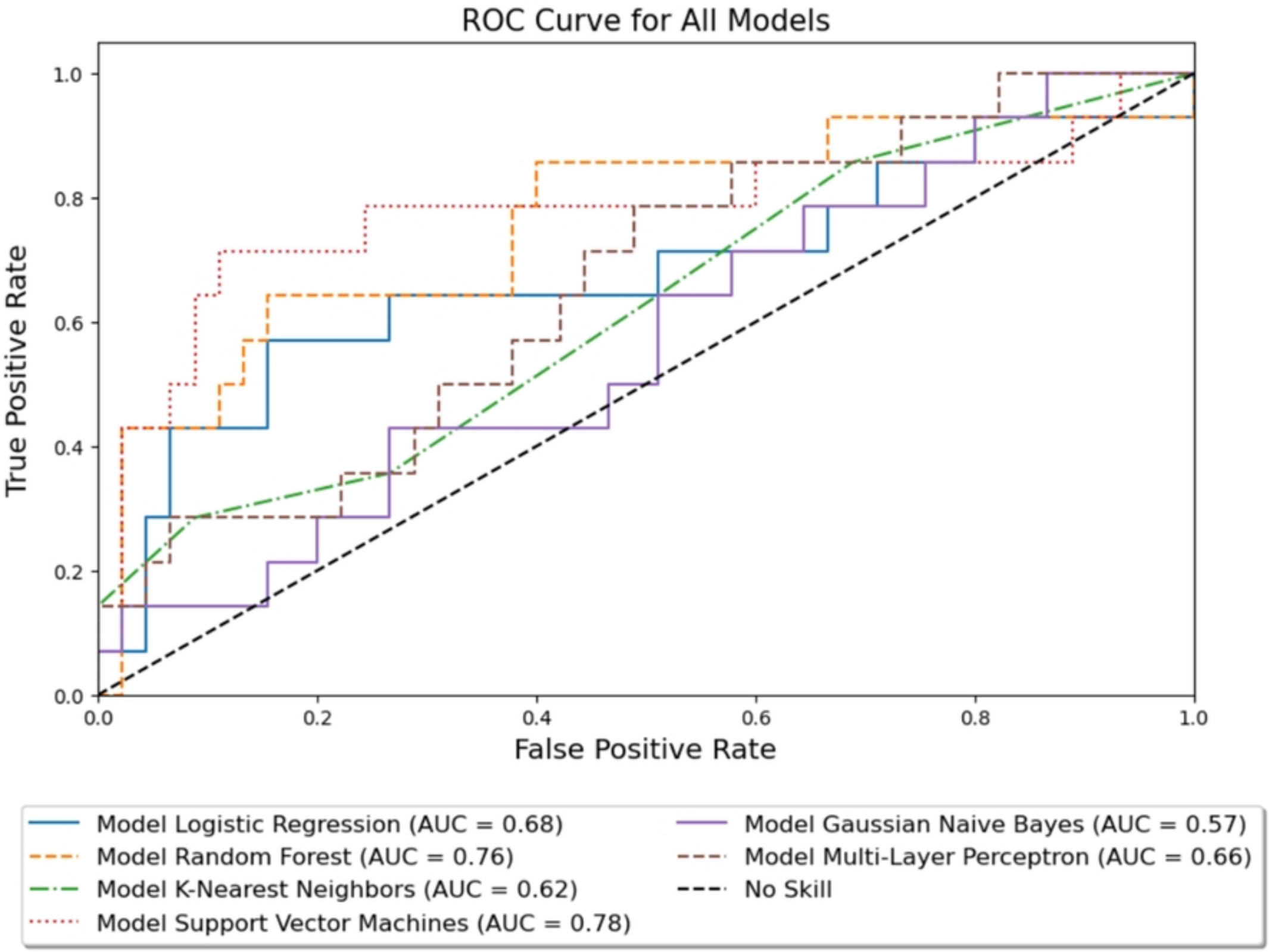
ROC Curve for All Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Model Logistic Regression (AUC = 0.68)
Model Random Forest (AUC = 0.76)
Model K-Nearest Neighbors (AUC = 0.62)
Model Support Vector Machines (AUC = 0.78)
Model Gaussian Naive Bayes (AUC = 0.57)
Model Multi-Layer Perceptron (AUC = 0.66)
No Skill

Figure 3