

1 **Advancing patient care: Machine learning models for predicting grade**  
2 **3+ toxicities in gynecologic cancer patients treated with HDR**  
3 **brachytherapy**

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

8 <sup>1</sup>Department of Radiation Oncology, Brown Cancer Center, University of Louisville School  
9 of Medicine, Louisville, KY, United States of America

10 <sup>2</sup>Department of Radiation Oncology, CARTI Cancer Center, Little Rock, AR, United  
11 States of America

12 <sup>3</sup>Department of Gynecologic Oncology, Brown Cancer Center, University of Louisville  
13 School of Medicine, Louisville, KY, United States of America

14 <sup>4</sup>Department of Bioinformatics & Biostatistics, University of Louisville School of Public  
15 Health and Information Sciences, Louisville, KY, United States of America

16 <sup>5</sup>Department of Radiation Oncology, Vanderbilt University Medical Center, Nashville, TN,  
17 United States of America

18

19 \* Corresponding author

20 E-mail: [aportocarrerob@pucp.edu.pe](mailto:aportocarrerob@pucp.edu.pe)

21

## 22 **Abstract**

### 23 **Background:**

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

### 28 **Objective:**

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

### 32 **Methods:**

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

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

#### 44 **Results:**

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

#### 52 **Conclusions:**

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

## 56 **Introduction**

57 Gynecological cancers rank among the most diagnosed malignancies affecting women  
58 on a global scale [1]. In the United States of America, it is estimated that there will be  
59 116,930 new cases and 36,250 deaths in 2024 attributable to gynecologic malignancies  
60 [2]. Treatments for gynecologic cancers include surgery, chemotherapy, and/or  
61 radiotherapy [3]. Brachytherapy is necessary in the management of locally advanced  
62 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

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

## 93 **Methods**

### 94 **Data Collection**

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

106 and HDR treatments. Toxicities were classified according to the Common Terminology  
107 Criteria for Adverse Events (CTCAE) version 5.0 [16], and the integrity of the database  
108 was reviewed three times by both a physician and a medical physicist to ensure its  
109 accuracy and reliability. For treatment planning, the dosimetry goals as detailed in the  
110 EMBRACE trials and ASTRO Clinical Practice Guideline were followed [6,17]. All patients  
111 received EBRT and Brachytherapy. The process used to calculate the total EQD2 dose  
112 has been described in-detail in a previous work, and follows the procedure suggested by  
113 ICRU 89 [18-19]. This study was approved by our institutional review board (IRB  
114 22.0117).

## 115 **Statistical Analysis**

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

## 125 **Data Preprocessing**

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

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

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

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

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

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

## 167 **Evaluation of Machine Learning Models**



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

$$175 \quad Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1) \quad Precision = \frac{TP}{TP + FP}$$

$$176 \quad (2)$$

$$177 \quad Recall = \frac{TP}{TP + FN} \quad (3) \quad F1 \text{ Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$178 \quad MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (5)$$

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

190 considering performance across all possible decision thresholds of the model [27]. In this  
191 work, the reported F1, recall, and precision scores are calculated for the positive class  
192 (patients that present a toxicity). For the AUC-ROC curve, the baseline denotes a random  
193 classifier, manifesting as a diagonal line with an AUC-ROC value of 0.5. Conversely, the  
194 PR curve's baseline reflects a situation where all classifications are assumed to be  
195 positive, resulting in a horizontal line on the precision-recall plot; the position of this line  
196 on the Y-axis is contingent upon the characteristics of the data under consideration [28-  
197 29]. These prediction metrics are calculated and reported for the training (without the  
198 SVM-SMOTE generated synthetic samples used for data balancing) and withheld test  
199 data (with the missing data-imputed for both). For the purposes of this work the authors  
200 have considered the top ML models as the ones with the highest test data F1 score.  
201 Confidence intervals of 95% were calculated assuming a normal distribution, as justified  
202 by the Central Limit Theorem [30].

203 
$$PM \pm 1.96 \sqrt{\frac{1}{n} \cdot PM \cdot (1 - PM)}$$

## 204 **Sequential Feature Selection**

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

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

## 224 **Machine Learning Algorithms**

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

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

245

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

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

## 269 **Results**

270 The data included demographic and clinical data for  $n = 233$  patients, of which  $n = 56$   
271 (24%) had a grade 3 or higher toxicity. The demographic, treatment, and tumor-related  
272 data are shown in Table 1. Patients who experienced grade 3 or higher toxicity were found  
273 to have longer follow-up (median 12.4 months versus 3.8 months), more likely to have  
274 low or very high comorbidity scores and had significantly higher HR-CTV values (median  
275 50 cc versus 39 cc,  $p = 0.041$ ).

276

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

	Full Cohort		No Grade 3+ Toxicity		Grade 3+ Toxicity		p-value
	n=233	100%	n=177	76%	n=56	24%	
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]	<b>&lt; 0.001</b>
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							<b>0.014</b>
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]	<b>0.041</b>
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

279

280 Table 2 compares median dose coverage to the tumor (V100%, D50%, D90%, and

281 D98%) and the dose to the organs at risk (OARs) by toxicity status. Patients with toxicities

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

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

		Full Cohort N=233 (100%)		No Grade 3+ Toxicity N=177 (76%)		Grade 3+ Toxicity N=56 (23%)		p-value
		Median	IQR	Median	IQR	Median	IQR	
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]	<b>0.043</b>

288

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

305

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



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

312 **Table 3. Training and testing performance metrics for the considered machine**  
 313 **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

314  
 315 **Table 4. Most important features as selected by the Sequential Feature Selection**  
 316 **algorithm (where appropriate) and found optimal hyperparameters for the top 3**  
 317 **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	

## 318 Discussion

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

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

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

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

350 Supervised machine learning has been utilized to perform classification tasks in various  
351 areas of healthcare such as for predicting diagnosis and prognosis of COVID-19 patients,  
352 prediction of hospitalization due to heart disease, and outcome prediction of infectious  
353 diseases [43-45]. To the authors' knowledge, this is the first analysis using and comparing  
354 multiple models for predicting grade 3 or higher toxicities in gynecologic cancer patients  
355 treated with external beam radiation and HDR interstitial or T&O brachytherapy. Through  
356 March 2020, there were only 53 published studies on the use of machine learning to  
357 predict radiation-induced toxicities [46], and through September 2023, only 14 studies  
358 had been published on deep learning models to predict toxicities from radiation treatment  
359 [47].

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

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

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

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

396 Regarding the scoring metrics, our study showed that the support vector machine was  
397 the best model for predicting grade 3 toxicities, obtaining a training F1 score of 0.61,  
398 accuracy of 0.82, normMCC of 0.75, precision of 0.63, recall of 0.6, AUC-ROC of 0.87,  
399 and AUC-PR of 0.6; whereas for that same model, the test data obtained an F1 score of  
400 0.63, accuracy of 0.80, normMCC of 0.75, precision of 0.56, recall of 0.71, AUC-ROC of  
401 0.78, and AUC-PR of 0.65. In the withheld test data, out of all the patients that had a  
402 toxicity (n = 14), 71% were correctly predicted by the model (TP = 10); and out of all the  
403 predicted cases, 56% represented a true toxicity event and were not false positives (FP  
404 = 8). Given the high level of uncertainty in whether patients will develop toxicities, this  
405 may be viewed as an adequate performance. An important detail that must be considered  
406 is that the precision value is as important as the recall, since during normal clinical  
407 practice it is equally as important to avoid false positives as it is to detect true positive  
408 cases. In particular, a toxicity prediction model may suggest that the physician consider  
409 lowering the dose to certain OARs to prevent these high-grade radiation-related side  
410 effects; an algorithm with good recall but prone to predicting false positives may lead to  
411 reducing the dose for a patient not susceptible to developing toxicities. This reduction, in  
412 turn, may involve sacrificing a portion of tumor coverage, potentially decreasing tumor  
413 control. For this reason, the F1 score emerges as the optimal metric for evaluating the  
414 model's performance. In future investigations within this area, prioritizing either the recall  
415 or the precision score, which is not replaceable by specificity, could be explored. Notably,

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

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

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

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

## 442 **Conclusion**

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

452

## 453 **References**

- 454 1. Costa M, Lai C. Coordinated efforts to harmonise gynaecological cancer care. *Lancet*  
455 *Oncol.* 2022;23(8):971-972. doi:10.1016/S1470-2045(22)00386-2
- 456 2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin.*  
457 2024;74(1):12-49. doi:10.3322/caac.21830
- 458 3. Kehoe S. Treatments for gynaecological cancers. *Best Pract Res Clin Obstet Gynaecol.*  
459 2006;20(6):985-1000. doi:10.1016/j.bpobgyn.2006.06.006



- 460 4. Robin TP, Amini A, Schefter TE, et al. Disparities in standard of care treatment and  
461 associated survival decrement in patients with locally advanced cervical cancer. *Gynecol*  
462 *Oncol*. 2016;143:319-25. doi:10.1016/j.ygyno.2016.09.009
- 463 5. Colson-Fearon D, Han K, Roumeliotis MB, et al. Updated Trends in Cervical Cancer  
464 Brachytherapy Utilization and Disparities in the United States From 2004 to 2020. *Int J*  
465 *Radiat Oncol Biol Phys*. 2024;119:154-62. doi:10.1016/j.ijrobp.2023.11.036
- 466 6. Potter R, Tanderup K, Schmid MP, et al. MRI-guided adaptive brachytherapy in locally  
467 advanced cervical cancer (EMBRACE-I): a multicentre prospective cohort study. *Lancet*  
468 *Oncol*. 2021;22:538-47. doi:10.1016/S1470-2045(20)30753-1
- 469 7. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in  
470 healthcare epidemiology. *Clin Infect Dis*. 2018;66(1):149-153. doi:10.1093/cid/cix731
- 471 8. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*.  
472 2015;349(6245):255-260. doi:0.1126/science.aac4520
- 473 9. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for  
474 predicting heart failure readmission and mortality. *ESC Heart Fail*. 2021;8(1):106-115.  
475 doi:10.1002/ehf2.13073
- 476 10. Singal AG, Mukherjee A, Elmunzer BJ, et al. Machine learning algorithms outperform  
477 conventional regression models in predicting development of hepatocellular carcinoma.  
478 *Am J Gastroenterol*. 2013;108(11):1723-1730. doi:10.1038/ajg.2013.332
- 479 11. Chan K, Lee TW, Sample PA, et al. Comparison of machine learning and traditional  
480 classifiers in glaucoma diagnosis. *IEEE Trans Biomed Eng*. 2002;49(9):963-974.  
481 doi:10.1109/TBME.2002.802012
- 482 12. Rajula HSR, Verlato G, Manchia M, et al. Comparison of conventional statistical methods  
483 with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*.  
484 2020;56(9):455. doi:10.3390/medicina56090455

- 485 13. Ley C, Martin RK, Pareek A, et al. Machine learning and conventional statistics: making  
486 sense of the differences. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(3):753-757.  
487 doi:10.1007/s00167-022-06896-6
- 488 14. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and  
489 combining techniques. *Artif Intell Rev.* 2006;26:159-190. doi:10.1007/s10462-007-9052-3
- 490 15. Saravanan R, Sujatha P. A state of art techniques on machine learning algorithms: a  
491 perspective of supervised learning approaches in data classification. *2018 Second*  
492 *International Conference on Intelligent Computing and Control Systems (ICICCS).*  
493 2018;945-949. doi:10.1109/ICCONS.2018.8663155.
- 494 16. National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE).  
495 Available at:  
496 [https://ctep.cancer.gov/protocolDevelopment/electronic\\_applications/ctc.htm](https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm). Accessed  
497 April 16, 2024.
- 498 17. Chino J, Annunziata CM, Beriwal S, et al. Radiation therapy for cervical cancer: executive  
499 summary of an ASTRO clinical practice guideline. *Pract Radiat Oncol.* 2020;10(4):220-  
500 234. doi:10.1016/j.prro.2020.04.002
- 501 18. Portocarrero-Bonifaz A, Syed S, Kassel M, et al. Dosimetric and toxicity comparison  
502 between Syed-Neblett and Fletcher-Suit-Delclos Tandem and Ovoid applicators in high  
503 dose rate cervix cancer brachytherapy. *Brachytherapy.* 2024;23(4):397-406.  
504 doi:10.1016/j.brachy.2024.03.003
- 505 19. International Commission on Radiation Units and Measurements. ICRU report 89:  
506 Prescribing, recording, and reporting brachytherapy for cancer of the cervix. *J*  
507 *ICRU.* 2016;13(1-2). doi:10.1093/jicru/ndw028
- 508 20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J*  
509 *Mach Learn Res.* 2011;12:2825-2830.

- 510 21. Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in  
511 classification and prediction problems. *SIGKDD Explor.* 2001;3(1):27–32.  
512 doi:10.1145/507533.507538
- 513 22. Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data  
514 classification. *Int J Knowl Eng Soft Data Paradigms.* 2011; 3(1), 4-21. doi:  
515 10.1504/IJKESDP.2011.039875
- 516 23. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling  
517 technique. *J Artif Intell Res.* 2002;16:321-357. doi:10.1613/jair.953
- 518 24. He H, Bai Y, Garcia EA, et al. ADASYN: Adaptive synthetic sampling approach for  
519 imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE  
520 World Congress on Computational Intelligence).* 2008; 1322-1328.  
521 doi:10.1109/IJCNN.2008.4633969.
- 522 25. Vujović Ž. Classification model evaluation metrics. *Int J Adv Comput Sci Appl.*  
523 2021;12(6):599-606. doi:10.14569/IJACSA.2021.0120670
- 524 26. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC  
525 AUC as the standard metric for assessing binary classification. *BioData Min.* 2023;16(1):4.  
526 doi:10.1186/s13040-023-00322-4
- 527 27. Bradley AP. The use of the area under the ROC curve in the evaluation of machine  
528 learning algorithms. *Pattern Recognit.* 1997;30(7):1145-1159. doi:10.1016/S0031-  
529 3203(96)00142-2
- 530 28. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical  
531 diagnostic test evaluation. *Caspian J Intern Med.* 2013;4(2):627.
- 532 29. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot  
533 when evaluating binary classifiers on imbalanced datasets. *PLoS One.*  
534 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432

- 535 30. Raschka S. Model evaluation, model selection, and algorithm selection in machine  
536 learning. arXiv [Preprint]. 2018 [cited 2024 Aug 11]. Available from:  
537 <https://arxiv.org/abs/1811.12808>
- 538 31. Guyon I, Elisseeff A. An introduction to variable and feature selection. *JMLR*. 2003;3:  
539 1157-1182. doi:10.1162/153244303322753616
- 540 32. Zekić-Sušac M, Pfeifer S, Nataša Šarlija N. A Comparison of Machine Learning Methods  
541 in a High-Dimensional Classification Problem. *BSRJ*. 2014;5(3):82-96. doi:10.2478/bsrj-  
542 2014-0021
- 543 33. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot  
544 when evaluating binary classifiers on imbalanced datasets. *PLoS One*.  
545 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
- 546 34. Li J, Wang S, Morstatter F, et al. Feature selection: A data perspective. *ACM Comput*  
547 *Surv*. 2017;50(6):1-45. doi:10.1145/3136625
- 548 35. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. *Pattern*  
549 *Recognit Lett*. 1994;15(11):1119-1125. doi:10.1016/0167-8655(94)90127-9
- 550 36. Molina LC, Belanche L, Nebot À. Feature selection algorithms: A survey and experimental  
551 evaluation. *2002 IEEE International Conference on Data Mining, 2002. Proceedings*.  
552 2002; 306-313. doi:10.1109/ICDM.2002.1183917.
- 553 37. Prusty S, Patnaik S, Dash SK. SKCV: Stratified K-fold cross-validation on ML classifiers  
554 for predicting cervical cancer. *Front Nanotechnol*. 2022;4:972421.  
555 doi:10.3389/fnano.2022.972421
- 556 38. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-  
557 validation, bootstrap and systematic sampling for estimating the generalization  
558 performance of supervised learning. *J Anal Test*. 2018;2(3):249-262. doi:10.1007/s41664-  
559 018-0068-2

- 560 39. Lee LJ, Damato AL, Viswanathan AN. Clinical outcomes of high-dose-rate interstitial  
561 gynecologic brachytherapy using real-time CT guidance. *Brachytherapy*. 2013;12(4):303-  
562 310. doi:10.1016/j.brachy.2012.11.002
- 563 40. Mesko S, Swamy U, Park SJ, et al. Early clinical outcomes of ultrasound-guided CT-  
564 planned high-dose-rate interstitial brachytherapy for primary locally advanced cervical  
565 cancer. *Brachytherapy*. 2015;14(5):626-632. doi:10.1016/j.brachy.2015.04.006
- 566 41. Mazon R, Fokdal LU, Kirchheiner K, et al. Dose-volume effect relationships for late rectal  
567 morbidity in patients treated with chemoradiation and MRI-guided adaptive brachytherapy  
568 for locally advanced cervical cancer: Results from the prospective multicenter EMBRACE  
569 study. *Radiother Oncol*. 2016;120:412-419. doi:10.1016/j.radonc.2016.06.006
- 570 42. Li R, Shinde A, Liu A, et al. Machine learning–based interpretation and visualization of  
571 nonlinear interactions in prostate cancer survival. *JCO Clin Cancer Inform*. 2020;4:637-  
572 646. doi:10.1200/CCI.20.00002
- 573 43. Muhammad LJ, Algehyne EA, Usman SS, et al. Supervised machine learning models for  
574 prediction of COVID-19 infection using epidemiology dataset. *SN Comput Sci*. 2021;2:1-  
575 13. doi:10.1007/s42979-020-00394-7
- 576 44. Dai W, Brisimi TS, Adams WG, et al. Prediction of hospitalization due to heart diseases  
577 by supervised learning methods. *Int J Med Inform*. 2015;84(3):189-197.  
578 doi:10.1016/j.ijmedinf.2014.10.002
- 579 45. Noorbakhsh-Sabet N, Zand R, Zhang Y, et al. Artificial intelligence transforms the future  
580 of health care. *Am J Med*. 2019;132(7):795-801. doi:10.1016/j.amjmed.2019.01.017
- 581 46. Isaksson LJ, Pepa M, Zaffaroni M, et al. Machine learning-based models for prediction of  
582 toxicity outcomes in radiotherapy. *Front Oncol*. 2020;10:790.  
583 doi:10.3389/fonc.2020.00790

- 584 47. Tan D, Nasir NFM, Manan HA, et al. Prediction of toxicity outcomes following radiotherapy  
585 using deep learning-based models: A systematic review. *Cancer Radiother.*  
586 2023;27(5):398-406. doi:10.1016/j.canrad.2023.05.001
- 587 48. Tian Z, Zhou Z, Shen C, et al. A machine-learning-based prediction model of fistula  
588 formation after interstitial brachytherapy for locally advanced gynecological malignancies.  
589 *Brachytherapy.* 2019;18:530–8. doi:10.1016/j.brachy.2019.04.004
- 590 49. Chen J, Chen H, Zhong Z, et al. Investigating rectal toxicity associated dosimetric features  
591 with deformable accumulated rectal surface dose maps for cervical cancer radiotherapy.  
592 *Radiat Oncol.* 2018;13:125. doi:10.1186/s13014-018-1068-0
- 593 50. Zhen X, Chen J, Zhong Z, et al. Deep convolutional neural network with transfer learning  
594 for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Phys Med*  
595 *Biol.* 2017;62:8246–63. doi:10.1088/1361-6560/aa8d09
- 596 51. Lucia F, Bourbonne V, Visvikis D, et al. Radiomics analysis of 3D dose distributions to  
597 predict toxicity of radiotherapy for cervical cancer. *J Pers Med.* 2021;11(5):398.  
598 doi:10.3390/jpm11050398
- 599 52. Cheon W, Han M, Jeong S, et al. Feature Importance Analysis of a Deep Learning Model  
600 for Predicting Late Bladder Toxicity Occurrence in Uterine Cervical Cancer Patients.  
601 *Cancers.* 2023;15(13):3463. doi:10.3390/cancers15133463
- 602 53. Peng Y, Nagata MH. An empirical overview of nonlinearity and overfitting in machine  
603 learning using COVID-19 data. *Chaos Soliton Fract.* 2020; 139.  
604 doi:10.1016/j.chaos.2020.110055
- 605 54. El Naqa I, Boone JM, Benedict SH, et al. AI in medical physics: guidelines for publication.  
606 *Med Phys.* 2021; 48(9):4711-4714. doi:10.1002/mp.15170
- 607 55. Ali MM, Pau BK, Ahmed K, et al. Heart disease prediction using supervised machine  
608 learning algorithms: Performance analysis and comparison. *Comput Biol Med.*  
609 2021;136:104672. doi:10.1016/j.combiomed.2021.104672

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

613

## 614 **Supporting Information**

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

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

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

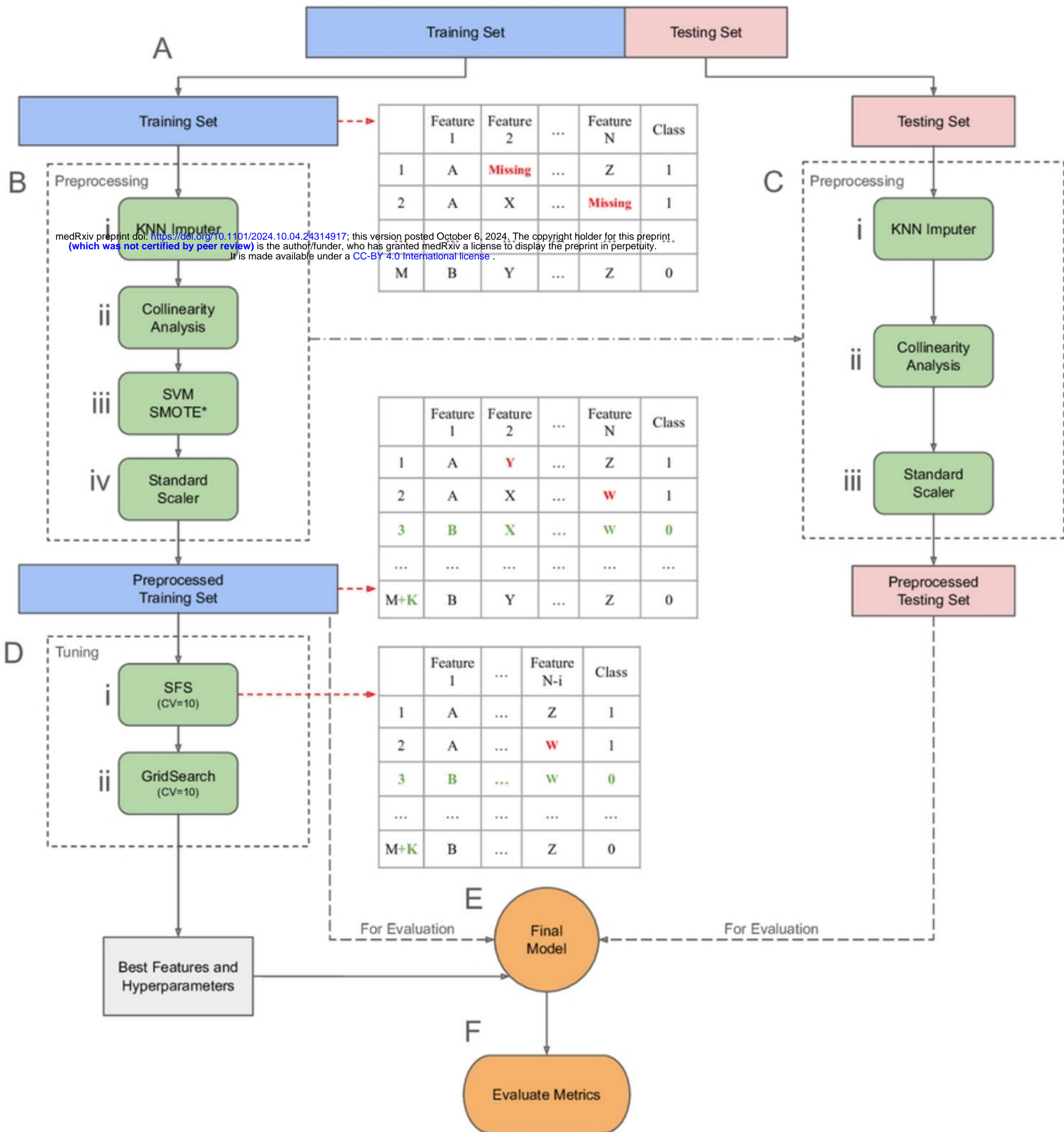


Figure 1



Precision-Recall Curve for All Models

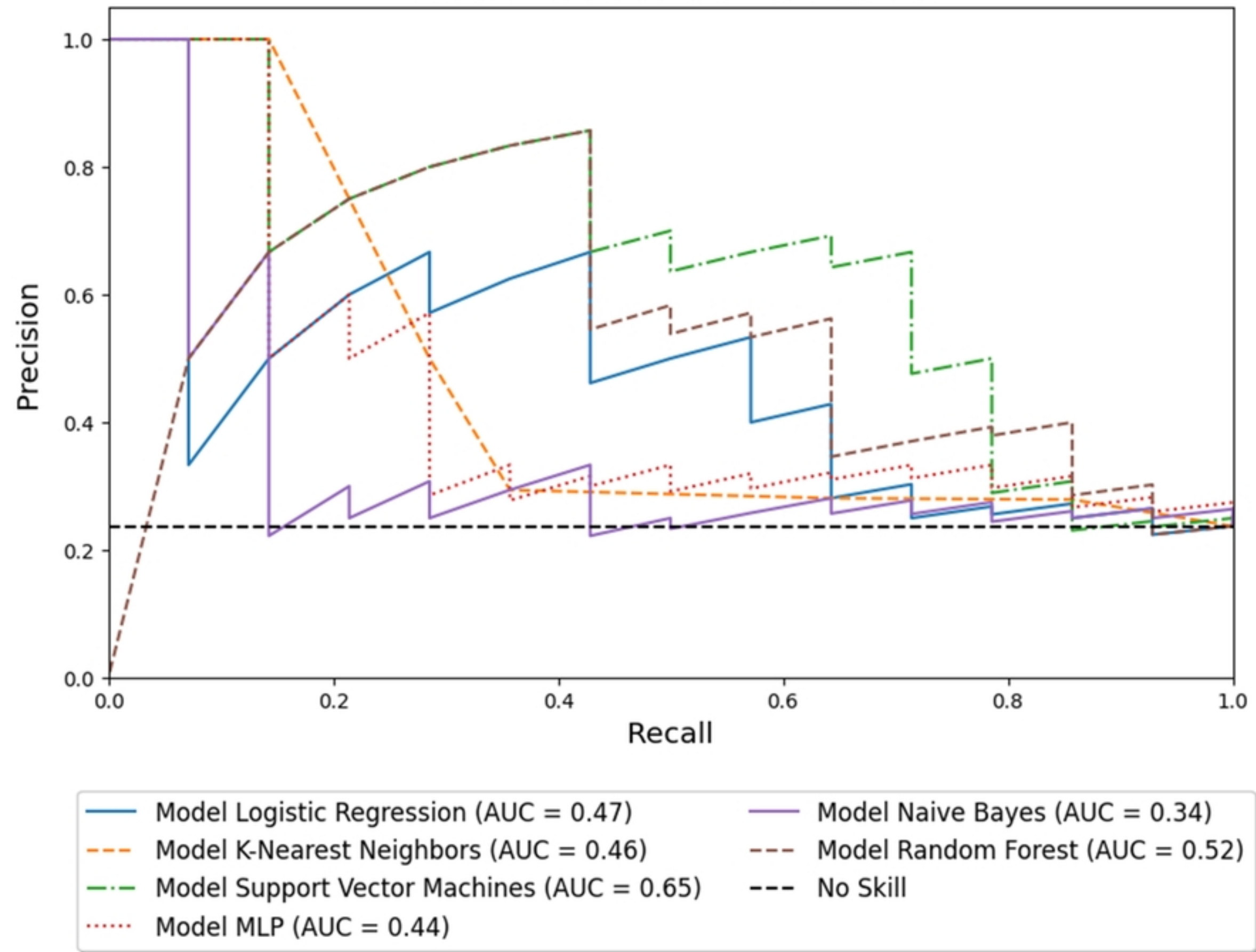
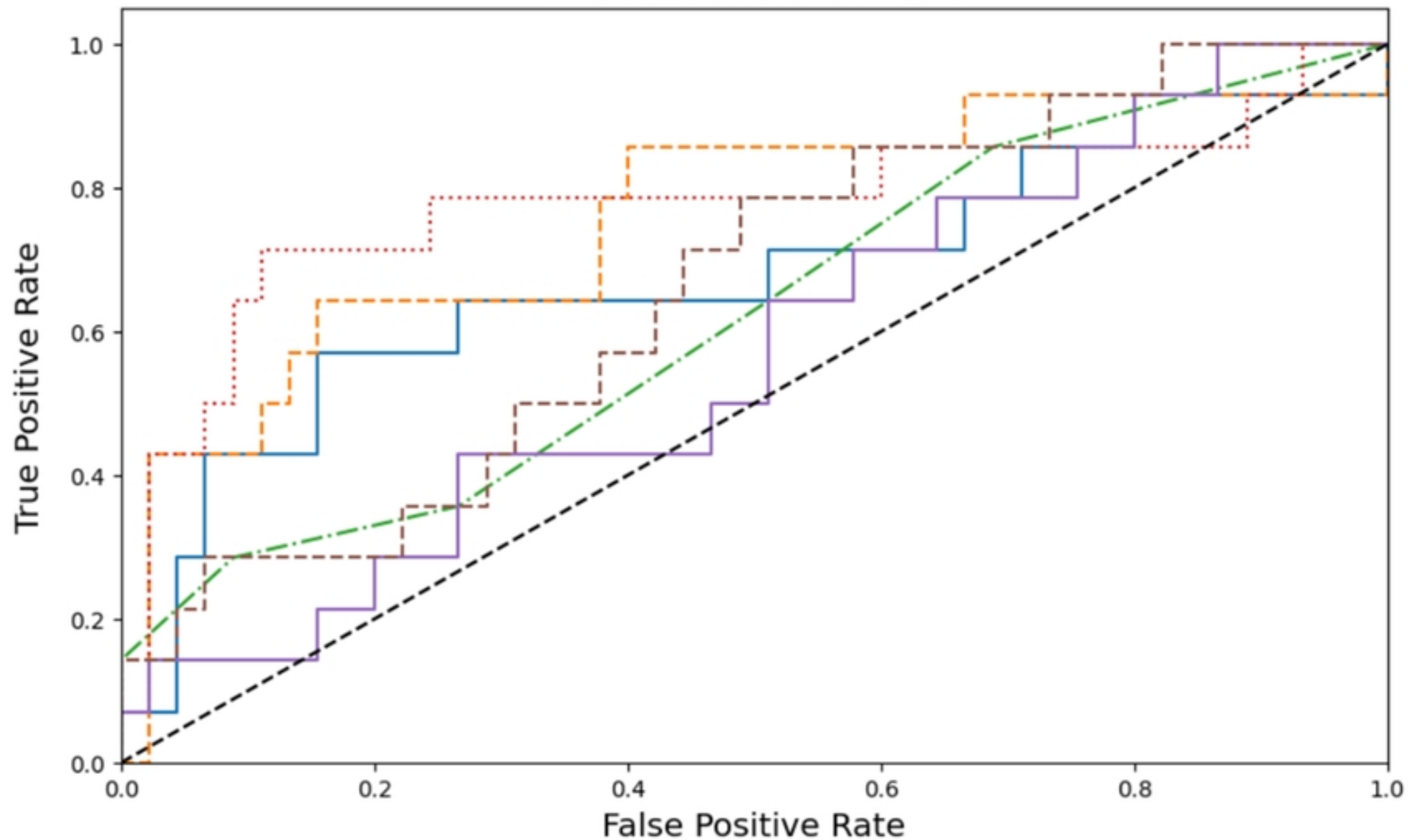


Figure 2

ROC Curve for All Models



- 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