ALMA MATER STUDIORUM UNIVERSITA' DI BOLOGNA

Scuola di Specializzazione in Medicina Nucleare

Direttore Prof. Stefano Fanti

The use of Machine Learning to predict 68Ga-PSMA-11 PET/CT result in different clinical settings of biochemical relapse after radical treatment for prostate cancer: comparison with a published nomogram.

TESI DI SPECIALIZZAZIONE

Presentata dal Dott. Relatore Chiar.mo Prof.

Francesca Serani Stefano Fanti

A.A. 2021/2022

Index

Abstract

Gallium-68 prostate-specific membrane antigen positron emission tomography (PSMA PET) is valuable for detecting prostate cancer. Its cost-effectiveness and long-term treatment outcomes based on PSMA PET results are uncertain. Correct patient selection for this imaging technique, especially in biochemical recurrence (BCR), is crucial for resource allocation and therapy planning. Machine learning (ML) algorithms show promise results in predicting outcomes, surpassing traditional nomograms. This study compares a published nomogram and a ML model for PSMA PET prediction in various BCR clinical settings. This comparison is essential to offer clinicians a comprehensive, practical, and highly accurate tool in the field of explainable Artificial Intelligence (exAI) for predicting PSMA PET result. We also explore whether ML, with a Logistic Regression with penalty equal to L1 (LR-L1) analysis can identify different variables, than the current ones used for the nomogram, to develop an alternative ML predictive model. The dataset included 703 prostate cancer patients with confirmed BCR post-radical therapy. Various ML algorithms were evaluated based on receiver-operating characteristic (ROC) curve AUC and accuracy. Logistic Regression emerged as the top-performing model (78.87% accuracy) compared to the nomogram (76.00%). LR-L1 analysis highlighted influential features, including PSA doubling time, PSA value at PSMA PET, clinical stage, and ISUP group, and the subsequent ML model developed showed an accuracy of 80.28%. The DeLong test did not show a statistical significance of the differences of AUC between the two ML models compared with the nomogram. Although Logistic Regression didn't significantly outperform the nomogram in PSMA positivity prediction, ML methods hold promise as complementary tools for medical applications, particularly in developing exAI for routine clinical use.

1 Introduction

Gallium-68 prostate-specific membrane antigen positron emission tomography (PSMA PET) is now a well-known and used imaging modality for the detection of prostate cancer, amongst professionals and patients. However, its cost-effectiveness in different clinical setting it is not yet established, and we still lack long term outcome analysis of the therapy based on PSMA PET imaging, even though it can change the management of treatment in a percentage of patients which goes from 30% to 50% [1,2], both in retrospective and in prospective context.
EAU guidelines have now included the use of PSMA PET both in the staging setting for high-risk prostate cancer patients and in the BCR context, however to date no outcome data exist to inform subsequent management [3], and clinicians have to be aware about the lack of long term outcome data of subsequent treatment changes.
Hence, PSMA PET imaging is not suitable for every patient, and its benefits may not apply universally, especially in cases of biochemical relapse (BCR) or biochemical persistence (BCP). Deciding when to employ this expensive method and basing clinical decisions on its results can lead to imprecise clinical decision-making. In the light of the advances of radionuclide imaging and therapy we need to be more attentive on how to allocate resources and understanding the implication of our choices. Therefore, new risk models for risk prediction are required. Machine learning algorithms are becoming more prevalent in creating predictive models because of their capacity to efficiently analyze large amount of data and big datasets. They have shown superior performance to clinical experts when it comes to predicting patient survival within a group of lung cancer patients [4]. Comparisons of outcome prediction models in other entities provided evidence that the reliability of machine-leaning based tools may be superior to those generated by traditional nomogram [5,6]. In particular it has been shown that machine learning model provides more personalized and reliable prognostic information than nomograms in patients with tongue cancer [5], suggesting that the combination of a nomogram - machine learning predictive model may help to improve tongue cancer management-related decisions.

1.1 PSMA PET: Cost effectiveness

At the moment of writing this paper the real cost-effectiveness of using PSMA PET in restaging after primary treatment it is not known. A recently published cost-effectiveness analysis, based on the use of PSMA in the staging setting, suggested that PSMA PET/CT was primarily

associated with increased costs in Europe and USA [7], in contrast from what was reported in the Australian setting [8] where PSMA PET/CT has lower direct comparative costs and greater accuracy compared to conventional imaging for initial staging of men with high-risk prostate cancer.

However, in Europe and USA, the use of PSMA PET/CT is appropriate from a health economic perspective as the costs for an accurate diagnosis using PSMA PET/CT seemed reasonably low compared to the potential consequential costs of an inaccurate diagnosis. This speculation however has to be verified by a prospective evaluation of patients at initial diagnosis [7].

The strive to correctly allocate resources is one of the many issues of modern medicine and it is a duty of the clinician to understand the clinical and economic implications of performing a costly examination.

1.2 What is Machine Learning (ML)?

In the field of Evidence-Based Medicine (EBM), the development of diagnostic and therapeutic protocols is based upon rigorous statistical analyses. These analyses employ inferential statistical methods to ensure that the conclusions derived from collected data have the possibility to be applied to the general population. The quality of these conclusions is directly proportional to the quantity and diversity of the data, as they enable more robust and precise decision-making.

Given the necessity to analyze larger quantity of data in the future in the medical field the application of machine learning may come in help and as stated in advance, it can outperform the usual statistical methods.

Machine learning (ML) is a field of study that involves using statistical learning and optimization methods that let computers analyze datasets and identify patterns [9], without being explicitly programmed. It is a subset of artificial intelligence (AI) that focuses on the development of algorithms and models that can learn from and make predictions or decisions based on data [10].

Machine learning algorithms can be used for decision processes to make predictions or classifications, evaluating the errors the model makes, when there are known examples, and through an “evaluate and optimize” process machine learning methodic can update itself until an accuracy threshold is reached [9].

There are many types of ML, and this is based on the presence or absence of human influence on raw data. These are divided into:

- **supervised learning:** the dataset being used has been pre-labeled and classified by users to allow the algorithm to see the accuracy of the performance.
- **unsupervised learning:** the raw dataset being used is unlabeled and an algorithm identifies patterns and relationships within the data without help from users.
- **semi-supervised learning:** the dataset contains structured and unstructured data, which guides the algorithm on its way to making independent conclusions. The combination of the two data types in one training dataset allows machine learning algorithms to learn to label unlabeled data.
- **reinforcement learning:** the dataset uses a "rewards/punishments" system, offering feedback to the algorithm to learn from its own experiences by trial and error [9].

Depending on the problem to solve and to the answer we want different algorithms can be chosen which are included in one of the categories above.

1.3 Categorization of problems for Machine Learning model development

Before taking to problem solving, the problem must be categorized suitably so that the most appropriate machine learning algorithm can be applied to it [11].
Thus, depending on the type of problem, an appropriate machine learning approach can be applied. Any problem in data science can be grouped in one of the following five categories:

• **Classification Problem** - A problem in which the output can be only one of a fixed number of output classes known apriori like Yes/No, True/False. Depending on the number of output classes, the problem can be a binary or multi-class classification problem.

• **Anomaly Detection Problem** - Problems that analyze a certain pattern and detect changes or anomalies in the pattern fall under this category. Such problems deal with finding out the outliers.
• **Regression Problem** - Regression algorithms are used to deal with problems with continuous and numeric output. These are usually used for problems that deal with questions like 'how much' or 'how many.'

• **Clustering Problem** - Clustering falls under the category of unsupervised learning algorithms. These algorithms try to learn structures within the data and attempt to make clusters

based on the similarity in the structure of the data. The different classes or clusters are then labeled. The algorithm, when trained, puts new unseen data in one of the clusters.

• **Reinforcement Problem** - Reinforcement algorithms are used when a decision is to be made based on past experiences of learning. The machine agent learns the behavior using trial and error sort of interaction with the continuously changing environment. It provides a way to program agents using the concept of rewards and penalties without specifying how the task is to be accomplished. Game playing programs and programs for temperature control are some popular examples using reinforcement learning [11].

2 Aims

Considered the development in the technological field we wanted to investigate whether ML could help in improving predictions. Therefore, we want to analyze the possible application of machine-learning based models for the correct decision-making whether to perform a PSMA PET in different clinical settings.

The first aim of this study is to compare the results of a previously published nomogram to predict PSMA PET positivity, in different clinical settings [12], with machine learning (ML) techniques and compare the performance of the two methodic.

The secondary aim is to find whether ML alone could identify an alternative predictive model and compare its performance with the published nomogram and the ML model identified.

To our best of knowledge, this is the first prediction tool using ML based method for predicting PSMA PET result.

3 Materials and Methods

3.1 Study Population

The cohort of patients included in this analysis was the same enrolled for the previous published nomogram [12]. The patients were enrolled through an open-label, prospective registry study performed in a single center institution (Prot. PSMA-PROSTATA; EudraCT: 2015-004589-27 OsSC). All patients provided signed informed consent prior to PSMA PET scan. Among all patients who received PSMA PET imaging in restaging setting (n = 1128, only patients who

had all clinical, pathological, imaging, and follow-up data available (n = 703) were included for analysis. All patients analyzed received PSMA PET at single referral center (Nuclear Medicine, University of Bologna, Italy) between March 2016 and November 2018 due to BCR (n = 627) or BCP (n = 76) after definitive therapy. BCR was defined as two consecutive PSA assays ≥ 0.2 ng/ml in patients treated with radical prostatectomy (RP) as primary treatment with/ without post-operative radiotherapy (RT) and as PSA ≥ 2 ng/ml above the nadir in patients treated with primary RT, in accordance with the Phoenix criteria [13] . BCP was defined as a PSA ≥ 0.1 ng/ml at 6 weeks after RP.

3.2 Clinical setting

Different clinical settings, as for the previous published nomogram [12], of PSA relapse were identified by referring physicians (urologist, radiation oncologist, and clinical oncologist) in a single-center multidisciplinary tumor board (Prostate Cancer Unit), and the overall population was grouped into 4 different categories of PSA recurrence, namely, first-time BCR (group 1, n = 325) defined as patients who had PSA nadir < 0.1 ng/ml after RP and subsequently experienced the first recurrence (group 1); PSA recurrence after salvage therapies (group 2, n = 241); BCP (group 3, n = 76); and advanced-stage PCa defined as patients with PSA progression under androgen deprivation therapy (ADT) (group 4, n = 61). No patients included in this analysis ever received any chemotherapy or ARTA.

68Ga-PSMA-11 was synthesized in the radiopharmacy of Nuclear Medicine, University-Hospital of Bologna, and prepared in a similar procedure as described by Eder et al. [14] and in previous publication [15,16]. A mean dose of 2 MBq/kg body weight (± 0.5 MBq/kg) of 68Ga-PSMA-11 was administered intravenously. 68Ga-PSMA-11-PET/CT was performed with a standard technique [17], as reported in our previous publication [15,16]. All studies were performed using a dedicated PET/CT state-of-the-art system (Discovery 690; Discovery MI, GE Healthcare, Milwaukee, WI, USA).

3.3 Images interpretation

All 68Ga-PSMA-11-PET/CT images were analyzed with dedicated software (eNTEGRA; GE Healthcare) and were independently interpreted with central reading by two nuclear medicine physicians with more than 8 years of experience in PET imaging (FC, PC). Readers were aware of all clinical data. In case of disagreement or undetermined lesion (event which occurred in 23

cases), a final diagnosis was reached by consensus considering the opinion of a third reader (SF) (majority rule in case of reader disagreement 2:1). Images were interpreted according to procedure guidelines [17,18], as reported in previous publication [15,16].

3.4 Nomogram

We used the nomogram constructed in another previously published study for evaluating PSMA PET result [12].

3.5 Machine learning training process

No formal sample size was elaborated. All patients with inclusion criteria were supposed to be eligible for the analysis and the number of participants was considered relevant for developing ML models.

The analysis was performed on Python software (Python Software Foundation, Oregon, USA) through the main libraries dedicated to Machine Learning development (SciPy, Scikit-learn, Pandas, NumPy, PyCaret, Matplotlib), and the variables tested were the ones which were included in the nomogram by the multivariable regression analysis (ISUP Group, PSA at PSMA PET, PSA doubling time (PSAdT), On-going ADT at BCR, Time To BCR and Clinical stage of BCR) [12].

ISUP Group	1 = 0 2 = 1 3 = 2 4 = 3 5 = 4
PSA at PSMA PET	Continuous numeric variable
PSA doubling time	Continuous numeric variable

On-going ADT at BCR	0 = No 1 = Yes
Time To BCR	0 = greater than 12 months 1 = less than 12 months
Clinical stage of BCR	0 = first time BCR 1 = Post-salvage Therapy 2 = PSA persistence 3 = Progression before systemic therapy

Table 1: Categorization of Variables used in machine learning training.

All the variables' items were labeled with numeric values to reduce possible spelling errors and omissions in each variable (Table 1).

The database was divided into a training set (n= 632) and an external validation set (n=71) (see Table 2 for the set characteristics). The two populations (training set and external validation set) were made homologous by minimizing the distance between their distribution, validating it through the use of the Kolmogorov-Smirnov test and the t-test for independent populations comparison. The data used for external validation were not used during the training process.

For model selection, PyCaret library was used with a 10-fold cross-validation on its set of classification algorithms and was trained on the whole training data set. After training, the algorithms were evaluated for the performance metrics of interest (Accuracy and Area Under the Curve (AUC)). The algorithm that showed the best AUC value was selected for the fine-tuning phase of hyperparameters.

The nomogram and the selected ML model that showed the best AUC were compared using the external validation data, generating positive PET result prediction probability.

The comparison using the same the external validation set on both nomogram and selected ML model, was necessary to ensure that the predictive tool used in medicine is convenient and accurate.

The first comparison metric considered was the AUC, and subsequently, the maximum accuracy value identified at the probability cut-off above which to consider the PET result as positive. For each model the accuracy and probability cut-off are paired.

We also investigated features contributions to determine the relative importance of each feature in a dataset when building the predictive model with the best AUC.

3.6 Comparison of the performance of the machine learning with the nomogram

The overall performances of the nomogram and the fine-tuned ML model were evaluated in terms of AUC and accuracy. The statistically significant difference between the AUCs of two ROCs was analyzed using the DeLong test. A two-sided p-value of less than 0.05 was considered statistically significant.

This comparison has been made for the ML predictive model to be used in medicine where clinicians ask for practical, accurate, and explainable results produced by algorithms. This approach is supported by the study of Holzinger et al., where human and machine explanations were compared using the system causability scale (SCS) to enable explainable AI (exAI) [19]. ExAI deals with the implementation of transparency and traceability of statistical black-box machine learning methods, particularly deep learning (DL). In the last years has been expressed the to even go beyond ExAI, adding "causability" to make AI explainable [19].

3.7 Development of a new model

In order to answer to our secondary aim (finding whether ML alone could identify an alternative predictive model) a subset of the variables present in the whole original dataset of the published nomogram [12] was tested with a Logistic Regression analysis with penalty value equal to L1 (LR-L1). From the original dataset the following variables were excluded:

- Radical Therapy: 19/703 of the patients performed External Beam Radiation Therapy (EBRT) as primary treatment, whereas 684/703 were treated with Radical Prostatectomy (RP). Due to a bias in numerosity this variable was excluded;
- GS Score: we decided to use ISUP group instead of the GS score;
- PSA velocity: given the high correlation with PSAdT, PSA velocity was excluded, whereas PSAdT was included in the analysis;
- For the TNM classification (T, N, M stage and margin status), initial PSA and age: not all the information listed were present for the whole dataset of patients, so we decided to not include these variables in the analysis.

The LR-L1 analysis permits to understand if there are variables with no importance in classification ML models and it is the analogous to the least absolute shrinkage and selection operator analysis (LASSO) executed for regression ML algorithms development. After the application of the LR-L1 algorithm it was repeated what described in section 3.5 to find the best fitting model for these variables.

The comparison of the latter model has been made with the nomogram and the previously performed ML model as described in section 3.6.

4 Results

In this study, we adopted the same formulation as previously published [12] and the nomogram variables of the patients, divided for training dataset and external validation set, are summarized in Table 2.

	Training Set	External Validation Set	p-value
ISUP Group			0.74
1	60(9.49)	7(9.86)	
2	130(20.57)	13(18.31)	
3	202(31.96)	21(29.58)	
4	205(32.44)	27(38.03)	
5	35(5.54)	3(4.23)	
PSA at PSMA PET (ng/mL)	1.34(0.39-1.34)	1.3(0.38-1.24)	0.94
PSA doubling time	8.13(3.5-9.73)	6.95(4.0-9.15)	0.21
On-going ADT at BCR			0.92
No	528(83.54)	59(83.1)	
Yes	104(16.46)	12(16.9)	
Time To BCR (months)			0.64
>12	444(70.25)	48(67.61)	
≤12	188(29.75)	23(32.39)	
Clinical Stage of BCR			0.3
First recurrence after primary treatment	294(46.52)	31(43.66)	
Recurrence after salvage therapy	219(34.65)	22(30.99)	
PSA persistence after primary treatment	66(10.44)	10(14.08)	
PSA progression before systemic therapies	53(8.39)	8(11.27)	
PET RESULT			0.93
Negative	308(48.73)	35(49.3)	
Positive	324(51.27)	36(50.7)	

Table 2: Clinical characteristics of patients for the nomogram variables included in ML model divided for training set and external validation set.

For the comparison with the nomogram a total of 15 non-fine-tuned classification models was trained with the six nomogram variables. Table 3 presents a summary of the performance metrics for different ML models when trained with the training dataset. The model with the highest AUC on the training dataset was selected: for this parameter the Logistic Regression model outperformed all other ML algorithms with a seed = 123.The previously published nomogram AUC was 0.82 (95% CI = 0.79–0.85). At ROC analysis, the best cut-off value to reliably predict positive PSMA PET was 40.00% (Accuracy = 76.00%). The machine learning algorithm (Logistic Regression), when tested on the external validation set, showed an AUC of 0.85 (95% CI = 0.77-0.93) and an accuracy of 78.87% with a cut off-of 53.00% (Figure 1 for ROC of the Logistic Regression ML model).

	Model	Accuracy	AUC
lr	Logistic Regression	0.7354	0.8097
lda	Linear Discriminant Analysis	0.7401	0.7963
gbc	Gradient Boosting Classifier	0.7311	0.7915
rf	Random Forest Classifier	0.7265	0.7876
ada	Ada Boost Classifier	0.7243	0.7834
xgboost	Extreme Gradient Boosting	0.6834	0.7773
lightgbm	Light Gradient Boosting Machine	0.7128	0.7764
et	Extra Trees Classifier	0.6856	0.7606
knn	K Neighbors Classifier	0.7174	0.7596
qda	Quadratic Discriminant Analysis	0.6087	0.7448
nb	Naive Bayes	0.5952	0.7396
dt	Decision Tree Classifier	0.6879	0.6890

Table 3: performance metrics for different ML models when trained with the training dataset.

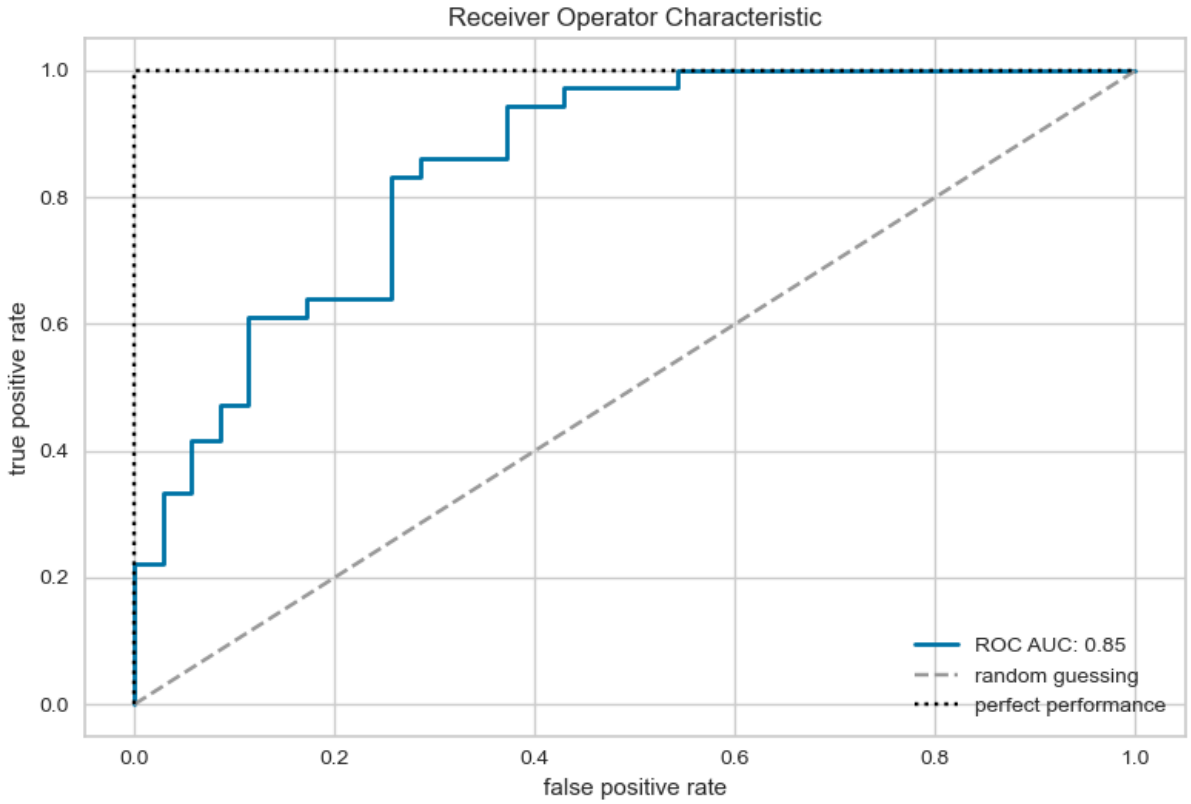

Figure 1: ROC of the external validation set for Logistic Regression

4.1 Evaluating input variables for importance

Logistic Regression model analyzes the importance of the variables with the permutation feature importance analysis, which measures the contribution of the various variable to the model.

Importance values of the variables for Logistic Regression model are presented in Table 4.

	Attribute	Importance
0	PSA at PSMA PET (ng/mL)	0.35
1	Clinical Stage of BCR	0.22
2	ISUP Group	0.19
3	On-going ADT at BCR	0.03
4	Time To BCR (months)	-0.03
5	PSA doubling time	-0.23

Table 4: Attributes (variables) importance for the model construction.

This analysis showed that the value of PSA at PSMA PET, the clinical stage of BCR, PSA doubling time and ISUP group of the surgery specimen had a greater influence on the model's performance to predict PSMA PET positivity. The predictors carry more weight in estimating the probability of PET positivity, when compared to time to relapse and on-going ADT at moment of PSMA PET. The latter two variables carry an importance value close to 0.00, both in negative and positive value.

4.2 Analysis for a new model

When the LR-L1 analysis was performed on the subset of the variables present in the whole original dataset of the published paper [12], as described in section 3.7, the only variables which were given importance were (in order of importance): PSA doubling time, PSA value at PSMA PET, clinical stage, and the ISUP group of the surgery specimen (Table 5). Time to relapse and Ongoing-ADT variables, which were included in the published nomogram showed no importance, could be removed from the prediction model.

	Attribute	Importance
0	PSA at PSMA PET (ng/mL)	0.36
1	ISUP Group	0.24
2	Clinical Stage of BCR	0.19
3	Pelvic LND	0.00
4	Adjuvant Radiotherapy	0.00
5	Adjuvant ADT	0.00
6	ADT during BCR	0.00
7	Salvage therapies	0.00
8	On-going ADT at BCR	0.00
9	Time To BCR (months)	0.00
10	PSA doubling time	-0.22

Table 5: LR-L1 analysis for variable selection for the development of a new model

After application of ML, on the reduced variables selected by the LR-L1 analysis, the best model was still the Logistic Regression model which showed an AUC of 0.85 (95% CI = 0.77-

0.93) when tested with external validation data, and an accuracy of 80.28% with a cut off set at 53.00%. Importance of the variables selected for the model created is summarized in Table 6.

	Attribute	Importance
0	PSA at PSMA PET (ng/mL)	0.41
1	Clinical Stage of BCR	0.25
2	ISUP Group	0.22
3	PSA doubling time	-0.24

Table 6: Importance of the variables for the model created with LR-L1 variable selection analysis.

4.3 Comparison of the performance of the nomogram with the Machine Marning algorithms

The nomogram exhibited an accuracy of 76.00% [12]. In our analysis the nomogram achieved the same accuracy of 76% when tested with the selected external validation data (Table 2). In contrast, the machine learning algorithm (Logistic Regression) demonstrated an accuracy of 78.87% when tested with the external validation data.

The ML model trained with variables selected with the LR-L1 analysis showed an accuracy of 80.28% (Table 7). The DeLong test did not show a statistical significance of the differences of AUC between the two ML models compared with the nomogram one (Figure 2 for the comparison of the ROC curves of the nomogram and the two ML models).

	Nomogram	Logistic Regression	Logistic Regression reduced dataset
True Positive	32	31	31
False Positive	13	10	9
True Negative	22	25	26
False Negative	4	5	5
AUC	0.82 95%CI(0.79-0.85)	0.85 95%CI(0.77-0.93)	0.85 95%CI(0.77-0.93)
Accuracy (cut-off)	76.00% (40.00%)	78.87% (53.00%)	80.28% (53.00%)
DeLong Value		0.33	0.34

Table 7: The performance metrics of the comparison between the nomogram and machine learning model on external validation set.

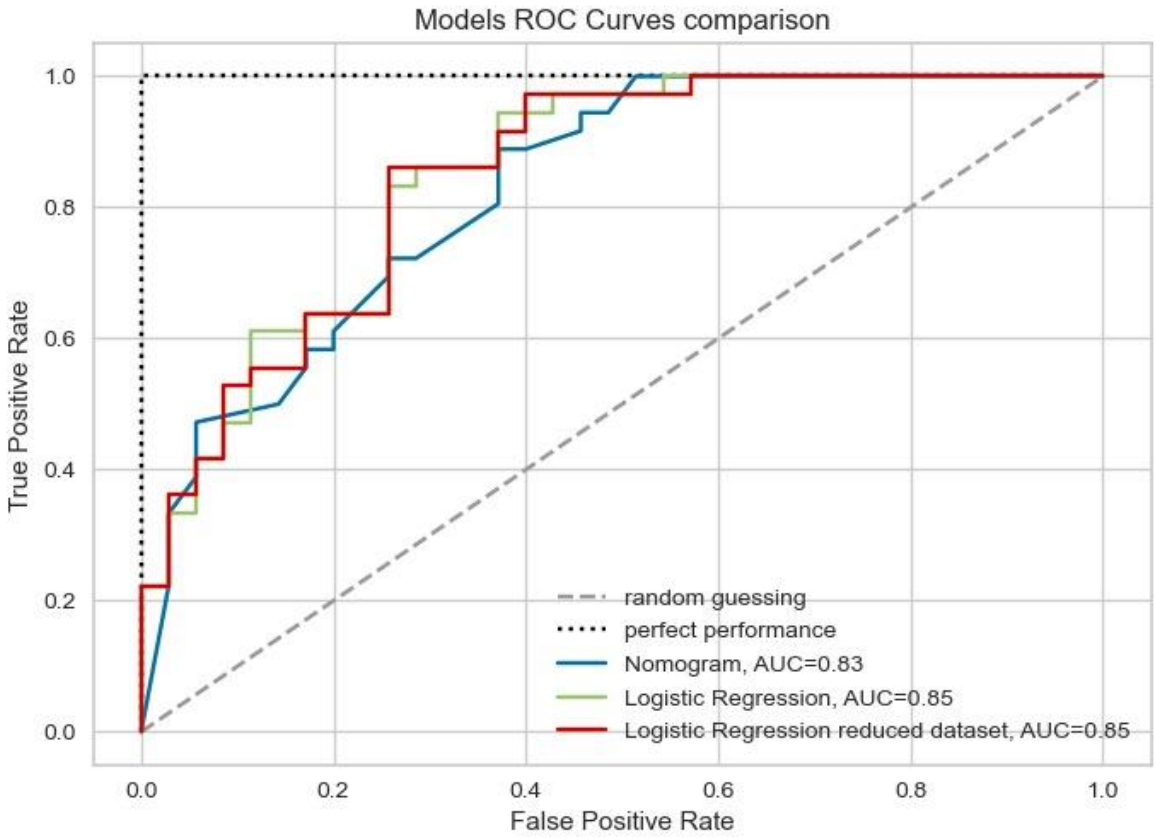

Figure 2: ROC of the nomogram, Logistic Regression ML with nomogram dataset and Logistic Regression for the LR-L1 selected dataset

5 Discussion

The medical community is constantly striving to develop and refine predictive tools, helping clinician to formulate the best decision for their patients, providing a more personalized care that tries to maximize patient outcomes while minimizing errors, diagnostic delays and being cost-effective. In modern evidence based medicine nomograms were and are constantly used in the medical field for the speculation of prognosis and future results and it is now emerging the great potential of the machine learning approach for dynamic prediction [20].

Even though ML and AI have now their place and their role in many real-life applications, such as voice recognition or spam e-mail detection, the application of these relatively new technologies in medical field is still unclear and it can be generally stated that it makes us worry about its supposed uncontrolled applications. With this study comparing a known nomogram with ML we wanted to investigate the possible integration of these new technologies in the medical field. Understanding to what extent the ML algorithms can be applied to make predictions in medicines, where nomograms are the best know prediction model, especially in the urology field [21,22].

With nomograms, which are complex regression analysis, made simple through graphical representation, the medical world has transitioned from a subjective predictive model rooted on personal experience to an objective model grounded in verifiable data. However, what has persisted unchanged is the subjective nature of interpreting this 'objective' model. Presently, a nomogram stands as the most precise model for forecasting a specific event's likelihood for an individual patient, drawing upon collected data from other patients [21,22].

With ML we tested various model and select the best one for our purpose, predicting PSMA PET result in different clinical settings of BCR after radical treatment for prostate cancer. The model selected, Logistic Regression, showed a slight superior performance than nomogram with AUC of 0.85 (95%CI = 0.77-0.93) when tested with external validation data, and an accuracy of 78.87% with a cut-off of 53,00%, compared to the nomogram ones where the AUC was 0.82 (95% CI = 0.79–0.85) with an accuracy of 76% for a cut-off value of 40%. However, this result was not statistically significant.

Nevertheless, this study gives interesting insights for the future application of ML models in medicine, especially for the application of what it is now called explainable Artificial Intelligence (exAI).

The importance of the variables (Table 4) showed that the most important variable for PSMA PET positivity prediction is the PSA value at the moment of PET/CT: this is an already known

an independent predictor for PSMA PET positivity showed not only by the nomogram, but also from other authors [15,23]. Both the two ML model showed this important feature.

In general, it can be stated that all the independent predictors of a positive PSMA PET were confirmed by the ML models, and that ADT at time of PSMA PET and time to recurrence could not be necessary for the clinical evaluation of a patient in which it has to be deemed whether to perform a PSMA PET or not.

One relevant feature to notice is that we should not be biased by the negative value of PSAdT: this testifies that for lower values of PSAdT greater is the probability of having a positive PSMA PET. This feature has already been largely investigated in literature and PSA kinetics it is now a known significant factor which influences PSMA PET result [23–25].

The interesting point to notice for ML models development in comparison with the know nomogram, is the difficult execution of a subgroup analyses for the different clinical setting. Whilst this could be performed with no concerns with traditional inferential statistics, performing a subgroup analysis in ML means creating a new model specific for that subgroup patients. From the result we have we can only say that based on the values of Table 1 for Clinical stage of BCR, the higher the value greater the probability of having a positive PSMA PET.

The ML algorithms cannot exactly tell the operator amongst the value for a given variable, which is the most important, even when performed in supervised setting, which means that the operator knows which kind of value and/or importance is giving to the variable. This comes intrinsic to the ML methodic which is a “black box” model [19]. The problem is that even if we understand the underlying mathematical principles and theories, such models lack an explicit declarative representation of knowledge [19].

When Logistic Regression analysis with penalty value equal to L1 was applied to the whole database, the ML algorithm was able to identify a different predictive model using less variables than the nomogram (namely: PSA doubling time, PSA value at PSMA PET, the clinical stage, and the ISUP grade of the surgery specimen) with an AUC of 0.85 (95% CI = 0.77-0.93) when tested with external validation data, and an accuracy of 80,28% with a cut off set at 53,00%: however, as stated earlier, the ML algorithm only gives us the answer, without giving us an explanation of it.

The machine learning model showed its ability to discern and comprehend relationships among input variables. The predictive accuracy exhibited by this model is particularly well-suited for medical applications, and the interest of the application of AI in the medical field, especially the oncologic one, can be dated since at least 15 years [26]. Despite the comparable predictive performance of the machine learning model with the previously published nomogram, it is

important to note that the nomograms provide a transparent method for estimating patient risk. For the ML model the only way to evaluate the performance for individual clinical decision-making is the use with an online risk calculator.

The transparency offered by the nomogram addresses concerns about the interpretability of results generated by ML models.

In the medical field the strive for an explanation of the results obtained from a clinical trial or any study is mandatory, in order for the results to be applied to the general population or as a justification of the money spent in a particular treatment or diagnostic method.

Therefore it becomes crucial to assess the causability (the quality of explanations) and explainability (why an algorithm/system produced a certain result) of the ML tools and in the last few years various attempt were made to measure the quality of these explanations: one example is the systematic causability scale (SCS) proposed by Holzinger et al. [27], which combines causability and explainability to achieve explainable medicine. This effort has been made due to the international concerns which are raised on ethical, legal and moral aspects of developments of AI in the last years, particularly in the medical domain [28].

For future studies, evaluating ML proposed diagnostic tool using the SCS assessment is important. The growing concerns about the human-AI relationship and the role of AI-based models in clinical decision support should be addressed with a solid understanding of causability and explainability [19,27]

5.1 Limitations

This study has certain limitations to consider. Both the nomogram and ML models were developed using retrospective cohorts, so in the future an analysis investigating with a greater, and hopefully prospective, population may be required to understand whether the ML model could significantly outperform the previous published nomogram. Furthermore, information on certain variables, such as TNM staging, a known intrinsic feature of tumor aggressiveness, was not available. Therefore, further calibration of the nomogram to include these variables and comparison with deep machine learning technologies would be worthwhile.

Another limitation of this study is that variables previously used as “categorized” variables (PSA at PSMA PET and PSAdT), for the ML model were used as continuous variables. This may be seen as a limitation in the direct comparison of the methodic, however, it may represent a strength for the ML algorithms application, which is able to understand the importance of the

trend in continuous variables, without the need of categorization in order to apply the multivariate analysis.
Despite these limitations, the findings of this study provide valuable insights into the possibility to integrate ML model when evaluating variables for predictive medical tools.

6 Conclusion

The Logistic Regression model demonstrated superior performance among the tested machine learning algorithms. While it did not significantly outperform the previously published nomogram in predicting PSMA positivity in prostate cancer patients, these findings imply that machine learning methods have potential value as complementary tools for implementation and comparison alongside established methodologies in the medical field. This holds particular relevance in the context of developing explainable AI for future routine clinical applications of these emerging technologies.

7 Bibliography

1. Habl G, Sauter K, Schiller K, Dewes S, Maurer T, Eiber M, et al. 68 Ga-PSMA-PET for radiation treatment planning in prostate cancer recurrences after surgery: Individualized medicine or new standard in salvage treatment. Prostate. 2017;77:920–7.

2. Bottke D, Miksch J, Thamm R, Krohn T, Bartkowiak D, Beer M, et al. Changes of Radiation Treatment Concept Based on 68Ga-PSMA-11-PET/CT in Early PSA-Recurrences After Radical Prostatectomy. Front Oncol. 2021;11:665304.

3. Mottet N, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer—2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. European Urology. 2021;79:243–62.

4. Bartholomai JA, Frieboes HB. Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. Proc IEEE Int Symp Signal Proc Inf Tech. 2018;2018:632–7.

5. Alabi RO, Mäkitie AA, Pirinen M, Elmusrati M, Leivo I, Almangush A. Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. Int J Med Inform. 2021;145:104313.

6. Lei H, Zhang M, Wu Z, Liu C, Li X, Zhou W, et al. Development and Validation of a Risk Prediction Model for Venous Thromboembolism in Lung Cancer Patients Using Machine Learning. Front Cardiovasc Med. 2022;9:845210.

7. Holzgreve A, Unterrainer M, Calais J, Adams T, Oprea-Lager DE, Goffin K, et al. Is PSMA PET/CT cost-effective for the primary staging in prostate cancer? First results for European countries and the USA based on the proPSMA trial. Eur J Nucl Med Mol Imaging [Internet]. 2023 [cited 2023 Sep 26]; Available from: https://doi.org/10.1007/s00259-023-06332-y

8. de Feria Cardet RE, Hofman MS, Segard T, Yim J, Williams S, Francis RJ, et al. Is Prostate-specific Membrane Antigen Positron Emission Tomography/Computed Tomography Imaging Cost-effective in Prostate Cancer: An Analysis Informed by the proPSMA Trial. European Urology. 2021;79:413–8.

9. 2uadmin. What Is Machine Learning (ML)? [Internet]. UCB-UMT. 2020 [cited 2023 Sep 27]. Available from: https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/

10. What is Machine Learning? | IBM [Internet]. [cited 2023 Sep 26]. Available from: https://www.ibm.com/topics/machine-learning

11. Alzubi J, Nayyar A, Kumar A. Machine Learning from Theory to Algorithms: An Overview. J Phys: Conf Ser. 2018;1142:012012.
12. Ceci F, Bianchi L, Borghesi M, Polverari G, Farolfi A, Briganti A, et al. Prediction nomogram for 68Ga-PSMA-11 PET/CT in different clinical settings of PSA failure after radical treatment for prostate cancer. Eur J Nucl Med Mol Imaging. 2020;47:136–46.
13. Abramowitz MC, Li T, Buyyounouski MK, Ross E, Uzzo RG, Pollack A, et al. The Phoenix definition of biochemical failure predicts for overall survival in patients with prostate cancer. Cancer. 2008;112:55–60.
14. Eder M, Neels O, Müller M, Bauder-Wüst U, Remde Y, Schäfer M, et al. Novel Preclinical and Radiopharmaceutical Aspects of [68Ga]Ga-PSMA-HBED-CC: A New PET Tracer for Imaging of Prostate Cancer. Pharmaceuticals (Basel). 2014;7:779–96.
15. Ceci F, Castellucci P, Graziani T, Farolfi A, Fonti C, Lodi F, et al. 68Ga-PSMA-11 PET/CT in recurrent prostate cancer: efficacy in different clinical stages of PSA failure after radical therapy. Eur J Nucl Med Mol Imaging. 2019;46:31–9.
16. Farolfi A, Ceci F, Castellucci P, Graziani T, Siepe G, Lambertini A, et al. 68Ga-PSMA-11 PET/CT in prostate cancer patients with biochemical recurrence after radical prostatectomy and PSA <0.5 ng/ml. Efficacy and impact on treatment strategy. Eur J Nucl Med Mol Imaging. 2019;46:11–9.
17. Fendler WP, Eiber M, Beheshti M, Bomanji J, Ceci F, Cho S, et al. 68Ga-PSMA PET/CT: Joint EANM and SNMMI procedure guideline for prostate cancer imaging: version 1.0. Eur J Nucl Med Mol Imaging. 2017;44:1014–24.
18. Fanti S, Minozzi S, Morigi JJ, Giesel F, Ceci F, Uprimny C, et al. Development of standardized image interpretation for 68Ga-PSMA PET/CT to detect prostate cancer recurrent lesions. Eur J Nucl Med Mol Imaging. 2017;44:1622–35.
19. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery. 2019;9:e1312.
20. Pickett KL, Suresh K, Campbell KR, Davis S, Juarez-Colunga E. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC Med Res Methodol. 2021;21:216.
21. Guillonneau B. Ceteris Paribus and Nomograms in Medicine. European Urology. 2007;52:1287–9.
22. Bianco FJ. Nomograms and Medicine. European Urology. 2006;50:884–6.

23. Eiber M, Maurer T, Souvatzoglou M, Beer AJ, Ruffani A, Haller B, et al. Evaluation of Hybrid ^{68}Ga-PSMA Ligand PET/CT in 248 Patients with Biochemical Recurrence After Radical Prostatectomy. J Nucl Med. 2015;56:668–74.
24. Afshar-Oromieh A, Avtzi E, Giesel FL, Holland-Letz T, Linhart HG, Eder M, et al. The diagnostic value of PET/CT imaging with the (68)Ga-labelled PSMA ligand HBED-CC in the diagnosis of recurrent prostate cancer. Eur J Nucl Med Mol Imaging. 2015;42:197–209.
25. Ceci F, Uprimny C, Nilica B, Geraldo L, Kendler D, Kroiss A, et al. 68Ga-PSMA PET/CT for restaging recurrent prostate cancer: which factors are associated with PET/CT detection rate? Eur J Nucl Med Mol Imaging. 2015;42:1284–94.
26. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007;2:59–77.
27. Holzinger A, Carrington A, Müller H. Measuring the Quality of Explanations: The System Causability Scale (SCS). Künstl Intell. 2020;34:193–8.
28. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25:1337–40.