SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Integration of Clinical, Genetic, and Molecular Features in Predicting Castration Resistance Events in Prostate Cancer: A Comprehensive Machine Learning Analysis

A. Mohammadi [1]***, M. Habibi*** [1]***, F. Parandin*** [2, *]

[1] *Department of Computer Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran.*

[2] *Department of Electrical Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran.*

Article Info

Article History:
Received 24 November 2023
Reviewed 11 January 2024
Revised 21 February 2024
Accepted 06 March 2024

Keywords:
Metastatic castration-sensitive prostate cancer
Castration resistance
Random forest classifier
Personalized medicine
Machine learning

[*] Corresponding Author's Email Address: *fa.parandin@iau.ac.ir*

Abstract

Background and Objectives: Metastatic castration-sensitive prostate cancer (mCSPC) represents a critical juncture in the management of prostate cancer, where the accurate prediction of the onset of castration resistance is paramount for guiding treatment decisions.

Methods: In this study, we underscore the power and efficiency of auto-ML models, specifically the Random Forest Classifier, for their low-code, user-friendly nature, making them a practical choice for complex tasks, to develop a predictive model for the occurrence of castration resistance events (CRE) . Utilizing a comprehensive dataset from MSK (Clin Cancer Res 2020), comprising clinical, genetic, and molecular features, we conducted a comprehensive analysis to discern patterns and correlations indicative of castration resistance. A random forest classifier was employed to harness the dataset's intrinsic interactions and construct a robust predictive model.

Results: We used over 18 algorithms to find the best model, and our results showed a significant achievement, with the developed model demonstrating an impressive accuracy of 75% in predicting castration resistance events. Furthermore, the analysis highlights the importance of specific features such as 'Fraction Genome Altered 'and the role of prostate specific antigen (PSA) in castration resistance prediction.

Conclusion: Corroborating these findings, recent studies emphasize the correlation between high 'Fraction Genome Altered' and resistance and the predictive power of elevated PSA levels in castration resistance. This highlights the power of machine learning in improving outcome predictions vital for prostate cancer treatment. This study deepens our insights into metastatic castration-sensitive prostate cancer and provides a practical tool for clinicians to shape treatment strategies and potentially enhance patient results.

Introduction

Prostate cancer accounts for a significant proportion of cancer cases, making up one in every five diagnoses. It is the most common cancer among men and the second leading cause of cancer-related deaths in men in the U.S. Metastatic prostate cancer, a more advanced stage, has shown an increasing incidence despite an overall decline in prostate cancer cases since 2000. Changes in prostate-specific antigen (PSA) screening recommendations in 2008 and 2011 have played a role in this trend. However, metastatic castration-sensitive prostate cancer (mCSPC) cases showed a significant 72% increase in 2013

Doi: 10.22061/jecei.2024.10406.697

compared to 2004.

This rise raises concerns because mCSPC is generally considered incurable and has a lower survival rate than localized prostate cancer. Although localized prostate cancer has a 100% 5-year survival rate, mCSPC's prognosis of mCSPC is less favorable, with a 5-year survival rate of 29.8%. Patients with de novo metastases and those whose cancer spreads after being initially diagnosed with localized disease may respond differently to treatment [1]-[5]. However, Prostate cancer is prevalent among men and a significant cause of cancer-related deaths in the Western world [6]. While androgen deprivation therapy (ADT) is commonly used to manage prostate cancer, approximately one-third of patients develop resistance, leading to castration-resistant prostate cancer (CRPC) [7]. Patients typically progress to CRPC within 18-48 months, with metastatic CRPC (mCRPC) being a major contributor to short median survival times [8], [9]. Docetaxel is the primary treatment for mCRPC, and studies have shown that combined treatment with prednisone significantly enhances quality of life and survival [10], [11]. However, despite its benefits, some patients become resistant to docetaxel therapy and discontinue treatment due to adverse events As docetaxel-based chemotherapy remains crucial for managing advanced prostate cancer, the ability to predict early discontinuation based on patient characteristics remains uncertain [12]-[15]. Metastatic castration-sensitive prostate cancer (mCSPC) is a type of prostate cancer that is initially responsive to androgen deprivation therapy (ADT) but eventually progresses to castration-resistant prostate cancer (CRPC) [16]. Castration resistance events (CRE) are defined as the development of CRPC, which is characterized by disease progression despite ADT [17]. CRE is a significant challenge in managing mCSPC, and identifying patients at a high risk of developing CRE is crucial for improving treatment outcomes.

In recent years, several studies have utilized machine learning algorithms, including RFC, to predict CRE in mCSPC. A retrospective study by Pan et al. in 2019 identified patients with rapid progression from hormone-sensitive to CRPC, using machine learning techniques [18]. Another study by Park et al. in 2020 aimed to predict CRPC in patients with prostate cancer using a machine learning approach [19]. Saito et al. constructed a new prognostic prediction model for patients with prostate cancer based on longitudinal data obtained from electronic health records using machine learning techniques [20]. These studies demonstrate the potential of machine learning algorithms, such as RFC, in predicting CRE in mCSPC, which can aid personalized treatment planning and improve patient outcomes. Certainly, our work stands out from previous studies in several ways. In this study, we addressed the prediction of CRE in mCSPC using the RFC algorithm. We extensively utilized diverse data, including clinical, genetic, and molecular features, to decipher specific patterns and correlations indicative of castration resistance. We thoroughly analyzed these patterns using the RFC algorithm and constructed a predictive model for CRE occurrence. This model not only exhibits a significant level of accuracy in predicting CRE but also serves as a practical tool for clinicians to inform treatment strategies and optimize patient outcomes. Our work advances the understanding of the mCSPC and provides a tangible solution for enhancing clinical decision-making in this domain.

The rest of the paper is organized as follows: Section 2 presents the Materials and Methods, detailing the patient background and the process of building the machine learning model. Section 3 discusses the Random Forest Classifier, its performance evaluation, and the impact of hyperparameter tuning. Section 4 delves into the analysis of ROC curves and AUC values in model evaluation. Section 5 highlights the importance of features in the predictive model. Section 6 discusses the findings from recent studies on the role of 'Fraction Genome Altered' and PSA level in mCSPC patients. Section 7 explores the transformation of clinical practice through data-driven approaches for prostate cancer management. Section 8 concludes the paper, summarizing the findings and their implications for the future of medical practice. Finally, Section 9 is devoted to the author contributions and compliance with Ethical Standards.

Materials and Methods Patient Background

The present study involves an examination of data from a cohort of 424 prostate cancer patients, focusing specifically on individuals with metastatic castration-susceptible disease. The focus was on patients with metastatic castration-susceptible disease at the time of the gene profile sampling. Comprehensive patient information, specimen details, treatment approaches, and clinical outcomes were obtained from a specialized clinical research database. The data were sourced from pathology reports, patient-reported smoking status, and complete medical records. The primary treating physician determined the castration resistance and metastasis, with potential biases carefully scrutinized by the research team. A dedicated research radiologist evaluated disease extent, location, and bone involvement using bone and CT scans. High-volume cases met specific criteria for visceral metastases or at least four bone metastases, whereas others were classified as having low-volume disease. The dataset encompassed a median age of 66 years (interquartile range, 59–72), including patients with both low-volume and high-volume de novo metastatic disease. The samples originated from the bone (14%), liver (4%), lung (6%), lymph node (22%), other soft tissue (6%), and prostate (48%). Castration-resistance events were recorded in 139 patients with metastatic castration-susceptible disease and 184 patients without castration-resistance events [20]. (Supplementary Table 1).

Predicting Castration Resistance in Cancer Using Machine Learning

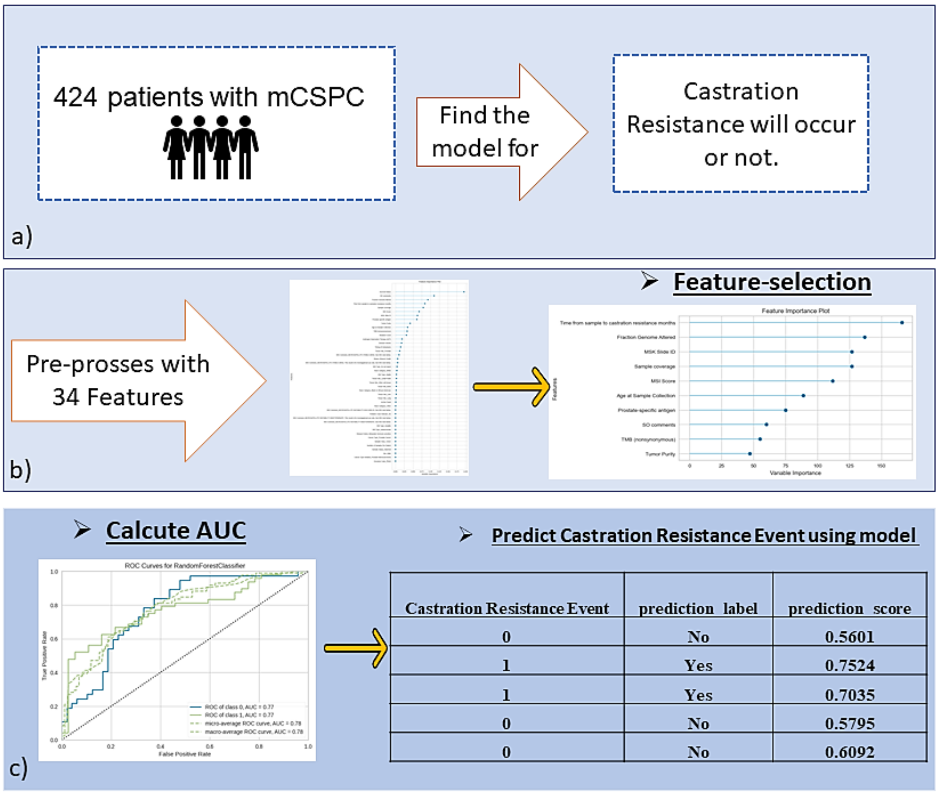
This diagram encapsulates three-phase approach for predicting Castration Resistance in mCSPC patients. Meticulous data collection, preprocessing, model selection, feature importance assessment, and model evaluation together represent a comprehensive journey towards achieving accurate predictions in a clinically significant context (Fig. 1).

In phase **a**), initially, a dataset containing information from 424 patients with mCSPC was collected.

Subsequently, these data were fed into a machine learning model, after which the best model was selected from among more than 30 different models, and a Random Forest classifier proved to be the most effective. Subsequently, during the ensuing phase, the model was fine-tuned using an optimization algorithm. The ultimate goal of this model was to predict the occurrence or absence of Castration Resistance based on 34 different patient features.

In phase **b**), important features are selected from among the various features during this stage

In phase **c**), The final model, configured according to these features, was evaluated. Ultimately, it was able to predict the occurrence of Castration Resistance with an accuracy of 75%, based on the importance of the features.


424 patients with mCSPC
Find the model for
Castration Resistance will occur or not.
a)
Pre-prosses with 34 Features
Feature-selection
Feature Importance Plot
Time from sample to castration resistance months
Fraction Genome Altered
MSK Slide ID
Sample coverage
MSI Score
Age at Sample Collection
Prostate-specific antigen
SO comments
TMB (nonsynonymous)
Tumor Purity
Features
0
25
50
75
100
125
150
Variable Importance
b)
Calcute AUC
ROC Curves for RandomForestClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.77
ROC of class 1, AUC = 0.77
micro-average ROC curve, AUC = 0.78
macro-average ROC curve, AUC = 0.78
Predict Castration Resistance Event using model
c)

Castration Resistance Event	prediction label	prediction score
0	No	0.5601
1	Yes	0.7524
1	Yes	0.7035
0	No	0.5795
0	No	0.6092

Fig. 1: Visualizing the journey.

From Data Preprocessing to Modeling: Building a Machine Learning Model

In this data preprocessing and modeling journey, a series of distinct stages was employed to prepare the data and construct a machine learning model. These stages encompass various transformations, addressing missing values, converting categorical variables, combining categorical information into target categories, and removing outliers. Commencing with raw data, which comprises unprocessed information, the process unfolds step-by-step. The Label Encoder is first utilized to convert categorical variables into numerical values.

The subsequent steps involve the application of SimpleImputer to handle missing values and iteratively refine results. The OrdinalEncoder and OneHotEncoder were then employed to convert categorical variables into ordinal or binary numerical values. The TargetEncoder stage leverages information from categorical variables as target categories to enhance the model performance. Subsequently, the RemoveOutliers stage eliminates outlier data points to improve the model performance. Finally, CleanColumn Names were employed for appropriate and standardized variable naming. Following this sequential order, the RandomForestClassifier algorithm was employed to create a machine-learning model tailored to the decision-making task. In this study, we utilized the Pycaret library, a powerful tool for automating machine learning workflows.

Table 1: Evaluation of the performance of various machine learning models

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT(Sec)
Random Forest Classifier	0.7374	0.8093	0.8058	0.7529	0.7738	0.4584	0.4682	1.003
Light Gradient Boosting Machine	0.7374	0.79	0.7692	0.7686	0.7651	0.4656	0.4709	0.689
Gradient Boosting Classifier	0.7348	0.7888	0.7853	0.7574	0.766	0.4571	0.4662	0.834
Extreme Gradient Boosting	0.72	0.7777	0.7282	0.7655	0.7415	0.4342	0.4391	0.804
Extra Trees Classifier	0.693	0.768	0.758	0.727	0.734	0.366	0.379	0.975
Linear Discriminant Analysis	0.6902	0.7537	0.7024	0.7387	0.7164	0.373	0.3775	0.518
Ada Boost Classifier	0.6844	0.7426	0.7384	0.7157	0.7227	0.3533	0.3581	0.768
Decision Tree Classifier	0.6609	0.6538	0.7018	0.6973	0.6953	0.3101	0.3148	0.46
Naive Bayes	0.6283	0.721	0.4658	0.7844	0.5665	0.2894	0.3256	0.92

The Random Forest Classifier

Machine learning algorithms have transcended the boundaries of engineering and optimization, permeating numerous other fields with their transformative capabilities [21]-[23].The Random Forest classifier has significantly affected the field of machine learning, particularly in predictive data mining. This ensemble learning technique uses labeled data samples to classify entities into distinct categories, train a model on one dataset, and evaluate its performance on an independent test dataset. It employs decision trees, which are fundamental tools in supervised learning, and enhances its performance by utilizing random sampling of data to create bootstrap samples and stochastic selection of input features for constructing individual decision trees. The power of the Random Forest technique lies in the combined strength of its decision-tree classifiers and their interaction, which significantly improves the classifier's ability to generalize beyond the training data. Comparative studies show that Random Forest accuracy rivals that of other ensemble techniques such as bagging and boosting. Its advantages include efficient operation on large databases, handling of multiple input variables without variable deletion, unbiased estimation of generalization error, effective handling of missing data, and robustness in maintaining accuracy even with significant data gaps. Additionally, its inherent parallel nature allows seamless implementation with cutting-edge technologies, such as multithreading, multicore processors, and parallel architectures [24]-[26].

This flowchart outlines the comprehensive pipeline for building, training, and evaluating machine learning models using a structured dataset. This process encompasses multiple stages, each contributing to the creation of an effective predictive model.

The objective of this study was to compare and evaluate the performance of various machine learning models for predicting a significant target variable. We used more than 18 algorithms to find the best model, and the best results are presented in the (Table 1). Interesting results were obtained, highlighting the crucial importance of selecting an appropriate model for prediction because of the valuable insights derived from this process. This is particularly vital for the design and implementation of data-driven decision-making systems (Fig. 2).

After carefully analyzing the results, the "Random Forest Classifier" was chosen as the preferred model due to its high accuracy (0.7374), substantial area under the ROC curve (AUC) value (0.8093), and commendable recall (0.8058). This model demonstrated exceptional predictive performance by finding a balance between the accuracy and recall. The F1 score (0.7738) also indicated its ability to effectively harmonize precision and recall.

Moreover, when considering other evaluation metrics such as the Kappa coefficient and Matthews Correlation Coefficient (MCC), the "Random Forest Classifier" showed significant values of 0.4584 and 0.4682, respectively. These results reinforced the selection of the "Random Forest classifier as the top-performing model for predicting the target variable. Furthermore, a comprehensive analysis highlighted that the "Light Gradient Boosting Machine" is an appealing choice. It exhibited a similar accuracy to the first model (0.7374), a reasonable AUC value (0.79), and acceptable recall (0.7692). Its F1 score (0.7651) indicated a well-balanced blend of precision and recall. Among the other models considered, the "Gradient Boosting Classifier" achieved notable results with an average accuracy (0.7348), satisfactory AUC value (0.7888), and recall (0.7853). This model demonstrated sensible equilibrium between accuracy and recall, as evidenced by its F1 score of 0.766. In conclusion, this analysis underscores the importance of selecting the correct model based on the significance of the evaluation metrics and the specific requirements of the problem at hand.

The "Random Forest Classifier" has been recommended as the top choice in this study for predicting the target variable, primarily due to its superior accuracy and overall performance (Table 1).

The table displays the performance metrics for the different machine-learning models used for classification. The metrics included Accuracy, Area under the Curve (AUC), Recall, Precision, F1 Score, Kappa, Matthews Correlation Coefficient (MCC), and Training Time (TT).

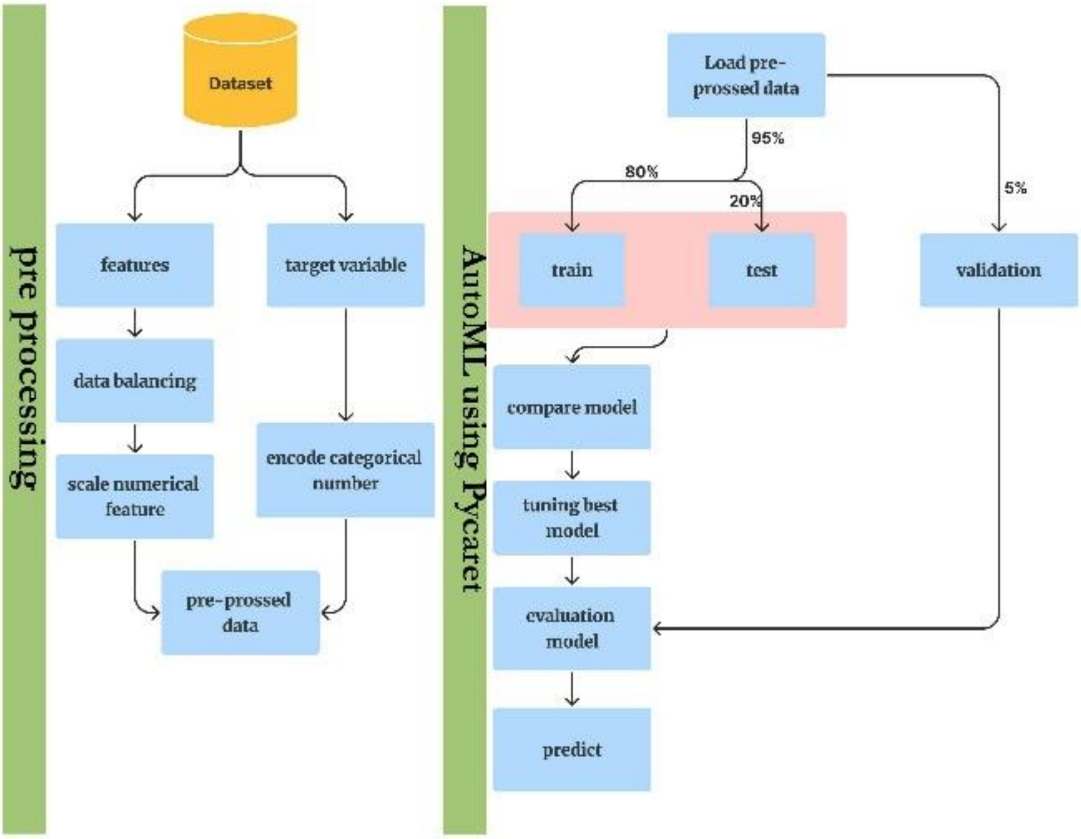


Fig. 2: flowchart of model.

Impact of Hyperparameter Tuning on Random Forest Classifier

Hyperparameter tuning plays a pivotal role in optimizing the performance of machine-learning models, enabling them to reach their full predictive potential. The Random Forest Classifier, a versatile ensemble learning algorithm, demonstrates the substantial influence of hyperparameter tuning on performance [27]-[31] . This study examines the impact of hyperparameter tuning on the performance of the Random Forest Classifier, drawing insights from the provided table that showcases various evaluation metrics (Supplementary Table 2).

The results showed varying performance across different splits. Accuracy ranged from 0.6471 to 0.8209, correctly categorizing the samples.

AUC values ranged from 0.7065 to 0.8757, indicating accurate class differentiation. Recall ranged from 0.6316 to 0.9231, highlighting the model's ability to identify positives. Precision, 0.6744–0.8250, managed the false positives well. F1 score, 0.6857–0.8571; balanced precision and recall. Kappa ranged from 0.2655 to 0.6322, suggesting good agreement, and MCC, 0.2675 to 0.6334, measured the prediction correlation. The average and standard deviation values summarize the performance metrics across the folds. The mean accuracy, AUC, recall, precision, F1 score, Kappa, and MCC were approximately 0.7377, 0.7936, 0.7649, 0.7712, 0.7642, 0.4663, and 0.4722, respectively. The standard deviations indicate the variations. In conclusion, this analysis highlighted the model's predictive probability across splits despite performance differences. Assessing key metrics revealed the model strengths and weaknesses, aiding real-world decision-making.

Analyzing the ROC Curves and AUC Values in Model Evaluation

Various metrics were used to assess the quality of the classification model. In this section, we analyze the ROC curves and Area Under the Curve (AUC) metric. The classification model in this study was trained to differentiate between two primary categories. The evaluation results (Fig. 3) of the test data are as follows.

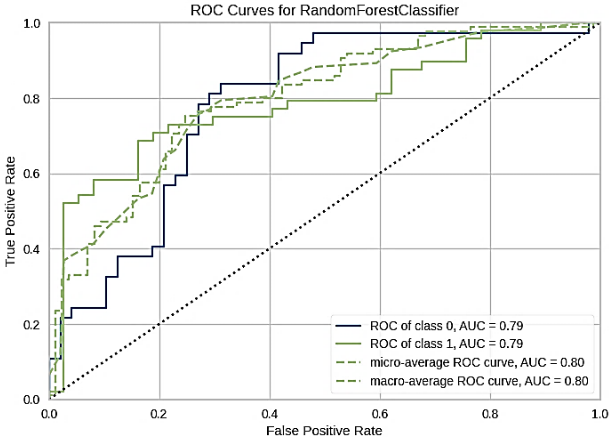


Fig. 3: ROC curves for random forest classifier.

The figure shows ROC curves for two classification models, class 0 and class 1, showing a balanced relationship between the True Positive Rate and False Positive Rate, indicating the effectiveness of the model in distinguishing between classes. A curve above this line indicates superior classification performance and informed decision making.

ROC curves serve as a powerful tool for evaluating a model's ability to distinguish between categories. Figure 4.1 displays the ROC curves for each class. As evident from the curves, the AUC for class 0 was 0.79, and that for class 1 was 0.79. These values indicate the capability of the model to discriminate between the two primary classes.

Micro-average and macro-average metrics were employed to assess the overall model performance. Here, the micro-average AUC calculates the mean of the true positive and false positive rates across all classes, providing a comprehensive assessment of the model's performance. This value was equal to 0.80, signifying the general ability of the model to distinguish between all classes. Based on the evaluation results, we conclude that the classification model exhibits high accuracy and proficiency in identifying different categories. The consistent AUC values across different classes, along with the micro-average and macro-average AUC values of 0.80, demonstrate the stability and reliability of the model's performance against test data.

Random Forest Classifier Performance Evaluation

The Random Forest classifier exhibited a satisfactory performance in the classification task. With an accuracy of 0.7529, the model showed the percentage of correctly classified instances out of the total. The Area Under the Curve (AUC) value of 0.7855 indicated the model's ability to discriminate between positive and negative classes, suggesting a promising overall discriminative capability (Table 2).

Table 2: Evaluating random forest classifier performance

Model	Accuracy	AUC	Recall	Prec	F1	Kappa	MCC
Random Forest	0.75	0.78	0.72	0.81	0.76	0.50	0.50

The table displays the performance metrics for the different machine-learning models used for classification. The metrics included Accuracy, Area under the Curve (AUC), Recall, Precision, F1 Score, Kappa, Matthews Correlation Coefficient (MCC), and Training Time (TT).

The confusion matrix provides a detailed overview of a machine-learning model's predictions versus actual outcomes, enabling a thorough evaluation of its performance. This matrix, with metrics such as accuracy and precision, highlights strengths and weaknesses, thereby guiding improvements for better predictions.
The confusion matrix provides further insight into the performance of the model (Fig. 4).

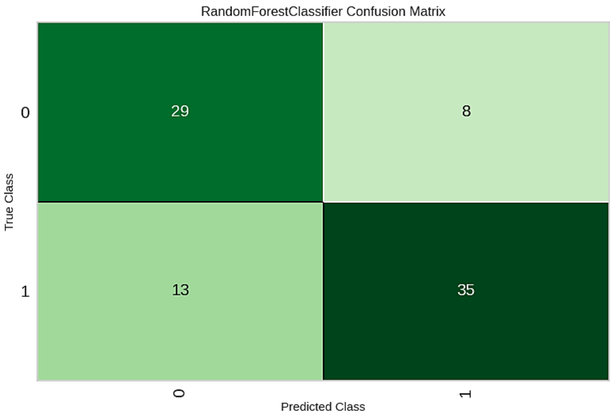


Fig. 4: Random Forest classifier confusion matrix.

The model achieved a recall score of 0.7292, reflecting its sensitivity for correctly identifying positive instances. At the same time, the precision score of 0.8140 highlights its ability to correctly identify positive instances from those that are predicted as positive. This balance between recall and precision is captured by the F1 score of 0.7692, which provides a harmonic mean and demonstrates the balanced performance of the model in handling both true positives and false negatives. A Kappa value of 0.5051 evaluates the model's agreement beyond chance with the actual outcomes, while a Matthews Correlation

Coefficient (MCC) of 0.5087 considers true positive, true negative, false positive, and false negative rates to provide a comprehensive classification measure.

In conclusion, the Random Forest classifier presents commendable accuracy and a balanced trade-off between precision and recall. The AUC value highlights the ability to discriminate between classes. However, further refinement could potentially enhance the predictive capacity and agreement with observed outcomes. This evaluation collectively provides valuable insights into the strengths and areas of the RF classifier for potential improvements in its classification capabilities.

The prediction result has been specified in the (Supplementary predict.csv) file.

In the file, the outcome of the prediction was specified using two elements: 'prediction_score' and 'prediction_label.' The term 'prediction_score' represents a numerical value that indicates the level of confidence or probability associated with the prediction. On the other hand, 'prediction_label' probably represents the assigned category or label that describes the prediction result. Together, these two pieces of information provide insights into the interpretation of the prediction."

Importance Feature

In descending order, the following features were ranked as the most important in our analysis. The significance and role of each of these features in the context of the subject are discussed below (Fig. 5).

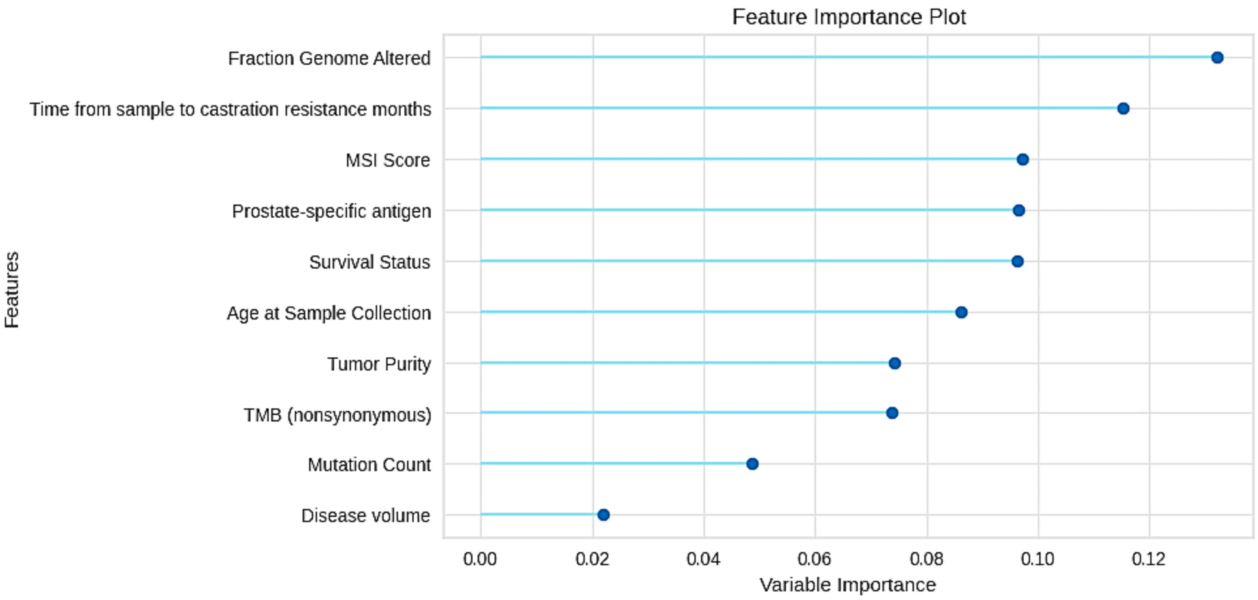


Fig. 5: feature importance plot.

This figure illustrates the results of the feature importance analysis on a tuned Random Forest classifier. This shows the varying impact of different features on the model's predictions, with higher importance scores indicating stronger influences. This insight emphasizes the significance of feature selection and tuning for optimizing the performance of the classifier.

Starting with the list of selected features, the 'Fraction Genome Altered' was introduced as the most critical feature. This feature indicates the proportion of the genome that has experienced alterations, and can serve as a representative of genomic fluctuations within the samples.

Next, the 'Time from Sample to Castration Resistance' is identified as another highly significant feature in this analysis.

The time taken for cancer to progress from the initial treatment to castration resistance offers essential insights into disease advancement and treatment responsiveness.

Following that, the 'MSI Score' is highlighted as a feature of great importance. This score reflects genetic instability in the tumor DNA, and an elevated MSI score might indicate a greater potential for responding to treatment through immunotherapy.

Prostate-specific Antigen (PSA) is deemed highly relevant as a biomarker for prostate cancer. Changes in PSA levels can provide vital information on disease progression.

Survival Status as Age at Sample Collection is another critical factor for predicting survival and understanding the impact of patient age on it. Understanding the influence of age on survival is crucial for determining disease prognosis.

Subsequently, Tumor Purity was mentioned as a determining feature of tumor cellularity within tissue samples. This metric can impact the accuracy of genetic analyses and treatment decisions.

The 'TMB (Nonsynonymous) Mutation Count' is highlighted as another significant feature in the order of importance. The number of non-synonymous mutations in the tumor genome indicates the extent of new genetic changes, potentially correlating with the immune therapy response.

Following that, the 'Mutation Count' signifies the overall population of mutations within the tumor genome and provides insights into genomic complexity.

This hierarchical analysis of the importance of selected features allows us to better understand the roles and impacts of each feature in predicting disease progression, patient prognosis, and treatment response. These selections can enhance the depth of the analysis and yield more comprehensive insights into the subject matter. All features are shown in (Supplementary Fig 1).

Discussion

Investigating the role of fraction genome altered and PSA level in (mCSPC) patient: Findings from a recent study

Based on the findings of this study (Singla et al., 2021), it can confidently be claimed that there is a strong correlation between the level of 'Fraction Genome Altered' and the occurrence of resistance. This study emphasizes that individuals with high genomic alterations during the initial treatment of prostate cancer are likely to face increased resistance to treatment and advanced castration. A precise analysis of the data from The Cancer Genome Atlas database clearly indicated that individuals with a higher number of altered genome regions were significantly associated with disease progression. This could serve as a robust indicator for predicting resistance and disease progression in prostate cancer. According to these investigations, it can be concluded that 'Fraction Genome Altered' could potentially serve as an indicator to identify individuals at higher risk of resistance to castration in prostate cancer. This unequivocally confirms that the results of this study will aid in enhancing diagnostic accuracy and suitable treatment planning for individuals with this type of cancer [32]. Based on research conducted by Nakanishi, Shotaro et al in 2021, the findings from this study revealed a strong correlation between PSA levels and castration resistance in patients with prostate cancer. This study emphasizes that an increase in PSA levels within three months of initiating treatment can serve as an independent indicator for progression towards castration-resistant prostate cancer and survival in patients diagnosed with metastatic hormone-sensitive prostate cancer. This research has clearly indicated that if the PSA level exceeds one percent of the pre-treatment value within the first three months of treatment, it can effectively predict the progression towards castration resistance and overall survival of patients.

The use of PSA levels as a significant factor in data analysis and treatment prediction for patients with prostate cancer has been substantiated. Furthermore, this study underscores that a more detailed examination of PSA levels within a specific time frame after treatment initiation can be a potent predictive tool for treatment outcomes. This study plays a pivotal role in enhancing diagnostic accuracy and refining treatment strategies for patients with prostate cancer [33].

Transforming clinical practice: Data-driven approaches for prostate cancer management

The study reveals the potential of machine learning algorithms in predicting castration resistance in metastatic castration-sensitive prostate cancer (mCSPC). By identifying predictive markers, the model can guide treatment strategies tailored to each patient's unique profile. The Random Forest Classifier achieved an impressive 75% accuracy in predicting Castration Resistance Event occurrence, demonstrating the potential of machine learning techniques in prostate cancer treatment. The study emphasizes the role of machine learning algorithms in enhancing personalized medicine. The synergy of clinical, genetic, and molecular features within the dataset has revealed patterns and correlations crucial in predicting castration resistance. However, further refinement and optimization are needed to enhance the predictive capacity and accuracy of the model. The interdisciplinary nature of the approach has broader implications for clinical practice, as the convergence of computer engineering, data analysis, and medical expertise could reshape clinical decisions. The study highlights the transformative potential of data-driven approaches in cancer research and personalized medicine, with the integration of advanced algorithms and interdisciplinary collaboration promising improvements in patient lives and shaping the future of medical practice.

Conclusion

In our extensive study aiming to predict Castration Resistance in Prostate Cancer using the Random Forest Classifier, we embarked on a meticulous exploration that involved evaluating over 18 algorithms to unearth the best predictive model. Our exhaustive efforts culminated in a significant achievement, with the developed model showcasing an impressive accuracy rate of 75% in predicting castration resistance events. The model's robust performance metrics, such as the AUC value of 0.7855 and a balanced F1 score of 0.7692, underscore its competence in accurately categorizing cases while maintaining precision and recall equilibrium. Further bolstering its classification capabilities, the Kappa coefficient (0.5051) and Matthews Correlation Coefficient (0.5087) validate its agreement with real-world outcomes. Additionally, our analysis of feature importance unveiled pivotal variables, with 'Fraction Genome Altered,' 'Time from Sample to Castration Resistance,' and 'MSI Score,' among others, providing crucial insights into disease complexity and patient prognosis. In conjunction with research by Singla et al. in 2021, our findings emphasize the significance of 'Fraction Genome Altered' and 'Prostate-specific Antigen (PSA)' as potent indicators for castration resistance. These insights, derived from our rigorous exploration across various algorithms, have the potential to significantly enhance diagnostic accuracy and refine treatment strategies for individuals confronting metastatic castration-sensitive prostate cancer.

Author Contributions

All authors designed, simulated, carried out the data analysis, collected the data interpreted the results, and wrote the manuscript.

Acknowledgment

This paper was written with the support of the Kermanshah Branch, Islamic Azad University.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Abbreviations

mCSPC	Metastatic Castration-Sensitive Prostate Cancer
CRE	Castration Resistance Event
ADT	Androgen Deprivation Therapy
CRPC	Castration-Resistant Prostate Cancer
PSA	Prostate-Specific Antigen

Data Availability

The data analyzed in this study were obtained from Metastatic Castration-Sensitive Prostate Cancer (MSK, Clin. Cancer Res. 2020) at http://www.cbioportal.org/study/clinicalData?id=prad_mcspc_mskcc_2020

Software and Code

The codes for our model are available at: https://github.com/alirezamohamadiam/Predicting_Castration_Resistance_Event

References

[1] Y. Kim, M. Alhassan, "Analyzing factors enabling prostate cancer screening behaviors among African American males in the south region using the Andersen's behavioral model of healthcare services utilization," J. Prev., 44(2): 253-266, 2022.

[2] H. Rahbar, P. Karabon, M. Menon, Q. D. Trinh, F. Abdollah, "trends in prostate-specific antigen screening since the implementation of the 2012 US preventive services task force recommendations," Eur. Urol. Focus., 4(6): 1002-1004, 2017.

[3] D. C. Parker, M. S. Cookson, "The changing landscape in the management of newly diagnosed castration sensitive metastatic prostate cancer," Investig. Clin. Urol., 61(1): S3-S7, 2020.

[4] W. K. B. Ranasinghe, N. A. Brooks, M. A. Elsheshtawi, J. W. Davis, T. K. Bathala, C. Tang, et al., "Patterns of metastases of prostatic ductal adenocarcinoma," Cancer, 126(16): 3667-3673, 2020.

[5] W. Ranasinghe, D. D. Shapiro, M. Zhang, T. Bathala, N. Navone, T. C. Thompson, et al. "Optimizing the diagnosis and management of ductal prostate cancer," Nat. Rev. Urol., 18(6): 337-358, 2021.

[6] E. Gupta, T. Guthrie, W. Tan, "Changing paradigms in management of metastatic castration-resistant prostate cancer(mCRPC)," BMC Urol., 14(55), 2014.

[7] D. Kristiyanto, K. E. Anderson, L. H. Hung, K. Y. Yeung, "Predicting discontinuation of docetaxel treatment for metastatic castration resistant prostate cancer (mCRPC) with randomforest," F1000Res., 5: 8353, 2016.

[8] M. S. Cookson, W. T. Lowrance, M. H. Murad, A. S. Kibel, "Castration-Resistant prostate cancer: AUA guideline amendment," J. Urol., 193(2): 491-499, 2015.

[9] A. Heidenreich, P. J. Bastian, J. Bellmunt, M. Bolla, S. Joniau, et al. "EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer," Eur. Urol., 65(2): 467-479, 2014.

[10] I. F. Tannock, R. de-Wit, W. R. Berry, J. Horti, A. Pluzanska, et al. "Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer," N. Engl. J. Med., 351: 1502-1512, 2004.

[11] D. P. Petrylak, "Docetaxel-based chemotherapy trials in androgen-independent prostate cancer: first demonstration of a survival benefit," Current Oncol. Rep., 7(3): 205-206, 2005.

[12] D. P. Petrylak, C. M. Tangen, M. H. A. Hussain, P. N. Lara Jr, J. A. Jones, M. E. Taplin, P. A. Burch, D. Berry, C. Moinpour, M. Kohli, et al., "Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer," N. Engl. J. Med., 351: 1513-1520, 2004.

[13] A. J. Templeton, F. E. Vera-Badillo, L. Wang, M. Attalla, P. De Gouveia, R. Leibowitz-Amit, J. J. Knox, M. Moore, S. S. Sridhar, A. M. Joshua, et al. "Translating clinical trials to clinical practice: Outcomes of men with metastatic castration-resistant prostate cancer treated with docetaxel and prednisone in and out of clinical trials," Ann. Oncol., 24: 2972-2977, 2013.

[14] C. J. Sweeney, Y. H. Chen, M. Carducci, G. Liu, D. F. Jarrard, M. Eisenberger, Y. N. Wong, N. Hahn, M. Kohli, M. M. Cooney, et al., "therapy in metastatic hormone-sensitive prostate cancer," N. Engl. J. Med., 373: 737-746, 2015.

[15] K. Deng, H. Li, Y. Guan, "Treatment stratification of patients with metastatic castration-resistant prostate cancer by machine learning," iScience, 23(2): 100804, 2020.

[16] D. Bansal, M. A. Reimers, E. M. Knoche, R. K. Pachynski," immunotherapy and immunotherapy combinations in metastatic castration-resistant prostate cancer," Cancers, 13(2): 334, 2021.

[17] L. H. Xiao, P. R. Chen, Z. P. Gou, Y. Z. Li, M. Li, L. C. Xiang, P. Feng, "Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen," Asian J. Andrology, 19(5):586-590, 2017.

[18] A. R. Hansen, et al., "Pantoprazole Affecting Docetaxel Resistance Pathways via Autophagy (PANDORA): Phase II Trial of high dose pantoprazole (Autophagy Inhibitor) with docetaxel in Metastatic Castration-Resistant Prostate Cancer (mCRPC)," The Oncologist, 24(9): 1188-1194, 2019.

[19] S. Saito, S. Sakamoto, K. Higuchi, et al., "Machine-learning predicts time-series prognosis factors in metastatic prostate cancer patients treated with androgen deprivation therapy," Sci. Rep., 13: 6325, 2023.

[20] K. H. Stopsack, et al. "Oncogenic genomic alterations, clinical phenotypes, and outcomes in metastatic castration-sensitive prostate cancer," Clin. Cancer Res., 26(13): 3230-3238, 2020.

[21] F. Parandin, A. Mohamadi, "Designing and optimizing a photonic crystal-based all-optical XOR gate using machine learning," Majlesi J. Electr. Eng., 18(1): 1-8, 2024.

[22] F. Parandin, M. R. Malmir, "Low delay time all optical NAND, XNOR and OR logic gates based on 2D photonic crystal structure," J. Electr. Comput. Eng. Innovations, 8(1): 1-8, 2020.

[23] M. Feli, F. Parandin, "A numerical optimization of an efficient double junction InGaN/CIGS solar cell," J. Electr. Comput. Eng. Innovations, 6(1): 53-58, 2018.

[24] V. Y. Kulkarni, P.K. Sinha, "Random forest classifiers: a survey and future research directions," Int. J. Adv. Comput., 36(1): 1144-1153, 2013.

[25] A. Parmar, R. Katariya, V. Patel, "A review on random forest: An ensemble classifier," in Proc. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018: 758-763, 2018.

[26] A. B. Shaik, S. Srinivasan, "A brief survey on random forest ensembles in classification model," in Proc. International Conference on Innovative Computing and Communications (ICICC), 2: 253-260, 2018.

[27] P. Probst, M. N. Wright, A. Boulesteix, "Hyperparameters and tuning strategies for random forest," WIREs: Data Min. Knowl. Discovery, 9(3): e1301, 2019.

[28] J. Hancock, T. M. Khoshgoftaar, "Impact of hyperparameter tuning in classifying highly imbalanced big data," in Proc. 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI): 348-354, 2021.

[29] H. L. Le, D. H. Tran, D. V. Chau, "A survey on the impact of hyperparameters on random forest performance using multiple accelerometer datasets," Int. J. Comput. Appl., 30(4): 351-361, 2023.

[30] K. E. Hoque, H. Aljamaan, "Impact of hyperparameter tuning on machine learning models in stock price forecasting," IEEE Access, 9: 163815-163830, 2021.

[31] H. J. P. Weerts, A. C. Mueller, J. Vanschoren, "Importance of tuning hyperparameters of machine learning algorithms," arXiv preprint arXiv:2007.07588, 2020.

[32] R. K. Singla, C. S. Sai, H. Chopra, S. Behzad, H. Bansal, R. Goyal, R. K. Gautam, C. Tsagkaris, S. Joon, S. Singla, B. Shen, "Natural products for the management of castration-resistant prostate cancer: Special focus on nanoparticles based studies," Front Cell Dev. Biol., 9: 745177, 2021.

[33] S. Nakanishi, M. Goya, et al., "Three-month early change in prostate-specific antigen levels as a predictive marker for overall survival during hormonal therapy for metastatic hormone-sensitive prostate cancer," BMC Res. Notes, 14: 227, 2021.

Biographies

Alireza Mohamadi received his Bachelor's degree in Computer Engineering from Islamic Azad University, Kermanshah Branch. He is particularly interested in applying Machine Learning to various areas, with a focus on Photonic Gates. He is passionate about exploring the potential of Artificial Intelligence, Machine Learning, and Deep Learning.

- Email: malireza718@gmail.com
- ORCID: 0009-0001-3698-481X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Maryam Habibi received her B.Sc. in Computer Software Engineering from Razi University in Kermanshah, Iran, in 2004. She then earned an M.Sc. in Artificial Intelligence from the Science and Research Branch of the Islamic Azad University in Tehran, Iran, in 2008. Currently, she is pursuing her Ph.D. at the same university with a primary focus on artificial intelligence. Since 2008, she has been a Lecturer at the Computer Engineering Department, Kermanshah Branch, Islamic Azad University in Kermanshah, Iran. She is a Faculty Member at the Computer Engineering Department, Kermanshah Branch, with research interests spanning artificial intelligence, machine learning, and computer vision.

- Email: ma.habibi@iau.ac.ir
- ORCID: 0000-0001-8939-8622
- Web of Science Researcher ID: NA
- Scopus Author ID: 15078505400
- Homepage: NA

Fariborz Parandin received the B.Sc. and M.Sc. degrees in Electrical Engineering from the University of Razi, Kermanshah, Iran, in 2000 and 2002, respectively. He obtained his Ph.D. degree in Optoelectronics from Razi University in 2017. He joined the Islamic Azad University, Kermanshah Branch, in 2008. He is currently the Associate Professor of Electrical Engineering at Islamic Azad University. His research interests include optoelectronics, semiconductor lasers, photonic crystals and applications of photonic integrated circuits.

- Email: fa.parandin@iau.ac.ir
- ORCID: 0000-0001-9044-3048
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
A. Mohammadi, M. Habibi, F. Parandin, "Integration of clinical, genetic, and molecular features in predicting castration resistance events in prostate cancer: A comprehensive machine learning analysis," J. Electr. Comput. Eng. Innovations, 12(2): 363-372, 2024.

DOI: 10.22061/jecei.2024.10406.697

URL: https://jecei.sru.ac.ir/article_2088.html