Egyptian Informatics Journal 33 (2026) 100895

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

Causal inference and explainable machine learning for analyzing treatment side effect in metastatic castration-resistant prostate cancer patients

Olutomilayo Olayemi Petinrin [a,b], Faisal Saeed [c,*], Hao Xue [d], Sumanta Basu [b], Shadi Basurra [c], Zhe Liu [a], Muhammad Toseef [a], Ibukun Omotayo Muyide [e], Ka-Chun Wong [a,f,*]

[a] Department of Computer Science, City University of Hong Kong, Kowloon Tong, 53025, Kowloon, Hong Kong Special Administrative Region
[b] Department of Statistics and Data Science, Cornell University, Ithaca, 14850, NY, USA
[c] DAAI Research Group, Department of Computing and Data Science, School of Computing and Digital Technology, Birmingham City University, Birmingham, B4 7XG, UK
[d] Department of Computational Biology, Cornell University, Ithaca, 14850, NY, USA
[e] College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA
[f] Shenzhen Research Institute, City University of Hong Kong, Shenzhen, 53025, Guangdong, China

ARTICLE INFO

Keywords:
Cancer metastasis
Causal inference
Explainable AI
Machine learning
Sensitivity analysis
Treatment effect

ABSTRACT

Optimal treatment recommendation for metastatic castration-resistant prostate cancer (mCRPC) are inherently diverse, being contingent upon individual patient response. Furthermore, treatment efficacy in specific patient cohorts can be influenced by confounding factors. Considering the substantial genetic heterogeneity among patients, generating population-level generalizations may compromise the precision and clinical applicability of predictive models. This study examines the prediction of treatment-induced adverse events in mCRPC patients using Explainable AI (XAI), focusing on both global and local levels of interpretability. Machine learning and other computational tools are often perceived as "black-box" techniques, largely due to the challenge of linking their internal processes to the final model outputs. Consequently, XAI offers crucial insight into the specific features that the algorithms prioritize for prediction, thereby illuminating the opacity and decision-making intricacies of these "black-box" models. Furthermore, causal inference was used to identify the attributes that specifically precipitate adverse events in patients with a smoking history. This analysis demonstrated that testosterone levels, prior analgesic use, and calcium levels act as confounders for adverse events within the smoking patients subgroup. The integration of causal inference and XAI establishes a robust and interpretable framework for making personalized treatment decisions in cancer care.

1. Introduction

Metastasis involves the movement of a tumor from a primary location to a secondary location. Metastatic cancer is the type of cancer that spreads from its original location to a distant part of the body, provided that there are conditions available that are necessary for its growth [1]. Glutathione (GSH), an antioxidant found in living organisms, can also promote tumor progression and increase metastasis when found in excess in the body [2]. Metastasis is the primary cause of death in cancer patients [3]. The 2025 estimated number of new cancer cases in the USA is 2,041,910, and the estimated number of fatalities is 618,120 [4]. There is a need to manage and control cancer to prevent metastases effectively. The detection of metastatic cancer is done based on the tumor cells. The tumor cells have features of the cells of the primary tumor location instead of the metastatic location. Examples include prostate cancer metastasizing in the bone [5] and colorectal cancer metastasizing in the brain [6]. In many cases, this is also referred to as Stage IV cancer [7].

A lot of research has been carried out on different metastatic cancer types based on their location. Some of these include prostate cancer [8,9], brain cancer [10], breast cancer [11], bone cancer [12]. Although some treatments are used for metastatic patients, these treatments are majorly used for controlling or suppressing the growth of the tumor, or palliative type of treatment that aids relief by improving quality of life. It is essential to choose the most suitable treatment according to the patient's response to treatment. Current research trends on metastasis involve the disruption of the enabling environment, which allows the spread of cancer cells [13], identifying and detecting the

* Corresponding authors.
E-mail addresses: faisal.saeed@bcu.ac.uk (F. Saeed), kc.w@cityu.edu.hk (K.-C. Wong).

https://doi.org/10.1016/j.eij.2026.100895
Received 15 January 2025; Received in revised form 6 December 2025; Accepted 18 January 2026
Available online 4 February 2026

specific genetic modifications in tumors [14], and building the immune system against cancer [15].

Due to the possible heterogeneity of cancer, the tumor might contain different cell collections with unique molecular signatures, which exhibit differences in sensitivity to treatment. This property of heterogeneity contributes to resistance and informs the need for accurate tumor heterogeneity assessment for the determination of corresponding effective therapy [16]. The available therapies and treatment for metastatic cancer often lead to tumor progression and treatment failure due to evolving resistance to treatment [17]. This often leads to degradation in the quality of life. Mechanisms related to specific metastatic cancer types can sometimes contribute to this treatment resistance [18]. Determining the optimal treatment recommendation is critical to maximize patient outcomes while minimizing the risk of adverse events and unnecessary healthcare costs.

Breakthrough in the application of machine learning, deep learning, and ensemble of methods has enabled the application of efficient solutions to problems in cancer research. Specifically, for metastasis, studies have been conducted where machine learning has been used for the prediction of survival and metastasis in patients [19–21], optimization of metastasis prediction methods [22,23], and comparison between methods [24–26]. The comparison by Wang et al. [26] reveals that the deep learning method used is not a significant improvement over classical machine learning methods and human specialists.

Causal inference, SHAP (SHapley Additive exPlanations) insight, and tree-based machine learning algorithms are powerful methods for predicting treatment outcomes and personalizing treatment decisions. Causal inference is a statistical framework that enables researchers to estimate the causal effect of an exposure or intervention on an outcome, while controlling for confounding factors. SHAP is a model-agnostic technique that provides interpretable explanations for the predictions of machine learning models, which can help clinicians and researchers understand the relative importance of different attributes in predicting patient outcomes. Tree-based machine learning algorithms, such as random forests and gradient boosting, are flexible and powerful methods that can handle complex interactions between attributes and can be used to develop predictive models for treatment outcomes.

In this study, we aim to predict the possibility of treatment-induced adverse event for mCRPC patients by using tree-based machine learning algorithms, SHAP explanation, and causal inference. Specifically, we will use electronic health records data of mCRPC patients who received docetaxel to train and validate several machine learning algorithms to predict treatment discontinuation. In addition, we will also use SHAP to interpret the importance of patient's history and other clinical attributes in predicting patient outcomes. Finally, we will estimate the causal effect of treatment on patient outcomes, including overall survival.

The integration of causal inference, SHAP, and tree-based machine learning algorithms can provide a comprehensive and interpretable approach for personalized treatment decision-making in cancer care. This study has the potential to inform clinical practice and improve outcomes for mCRPC patients by identifying the optimal treatment recommendation and facilitating personalized treatment decisions. In essence, risk stratification based on treatment continuation or discontinuation can be deduced.

2. Data description and visualization

The dataset used in this study was retrieved from Project Data Sphere [27]. It contains a record of metastatic castration resistant prostate cancer (mCRPC) patients in a Phase III clinical trials of docetaxel treatment. The dataset is a multicohort collection of three other datasets namely, Ascent-2, Celgene, and Venice dataset. Ascent-2 cohort is from Memorial Sloan Kettering Cancer Center. Celgene cohort is from Celgene, and Venice is from Sanofi. Models built with multicohort dataset provide more stability and robustness while maintaining superior performance compared to models built with single-cohort datasets [28]. The dataset contains the records of 1600 patients with 131 attributes. These attributes cut across the metastasis location, patient's lab values, demography, prior procedures, prior medications, and medical history (prior diagnosis and prior disease). For initial cleaning, we removed seven attributes, which are either identifiers, empty, or replicate of another column. Hence, we were left with 124 attributes. In addition, we filled some missing binary attribute cells as stipulated in the data dictionary. Regardless, some of the attributes have a high percentage of missing values as shown in Fig. 1.

In subsequent analyses, we handle the missing values using multiple imputation strategy. The dependent attribute which is the outcome to be determined is the "DISCONT" attribute. This indicates whether a patient should continue or discontinue the treatment.

3. Materials and methods

The hardware specification for all analyses are as follows: Windows 10 64-bit Operating System running on X64-based processor computer with 16 GB RAM. Processor specification is Intel (R) Core (TM) i7-10510U CPU @ 1.80 GHz. Analyses in this study were carried out using Microsoft Excel, Python, and R programming languages with several packages such as scikit-learn, shap, grf, and plotly. A flow diagram which depicts the workflow is shown in Fig. 2

3.1. Data preprocessing

For accurate analysis, the dataset was cleaned and preprocessed accordingly. The data dictionaries associated with the data provided the necessary guidelines for filling in some of the missing information. Attributes which do not contribute to prediction analysis, including patient's masked ID and cohort name, were removed. The attributes with missing values were filled using a multiple imputation method.

3.1.1. Multiple imputation by chained equations

For the imputation of other relevant missing values, instead of the single imputation methods such as mean, median and mode, we have used a multiple imputation method [29]. The multiple imputation method is used to address the problem of standard errors that arise from the use of traditional methods such as single imputation methods. By using multiple imputation method, a valid and unbiased estimate of missing data can be made from the available data based on association [30]. Also, compared to traditional methods used for handling with missing data, multiple imputation helps to retain information and deals with the issue of systematic dropout [31].

In this study, the specific multiple imputation method we used is a variant of Multivariate Imputation by Chained Equations (MICE) [32]. The variant used is the K-Nearest Neighbor (KNN) imputer, in which the values of the k-nearest neighbors of a sample with missing values is used to estimate the missing values for such samples. In this study, we have set our parameter k as 5, and the similarity metric as the Euclidean distance. The technique generates multiple imputed datasets before combining the datasets into a single estimation of the missing values.

3.1.2. Multicollinearity

The variance of coefficient estimates can increase due to multicollinearity. This increase in variance intensifies the sensitivity of the estimates to minor changes in the model. Multicollinearity occurs when there are correlations between several independent attributes in the data. The existence of multicollinearity among the independent attributes makes statistical inference from the model less reliable. Although multicollinearity might not affect the accuracy of a model, it can affect the interpretability and make it difficult to determine the effect of each attribute on the model. To detect attributes with multicollinearity, we use the Variance Inflation Factor (VIF), as explained in Appendix A. The importance of multicollinearity detection is attributed to the reduction it causes to the statistical significance of the independent attributes.

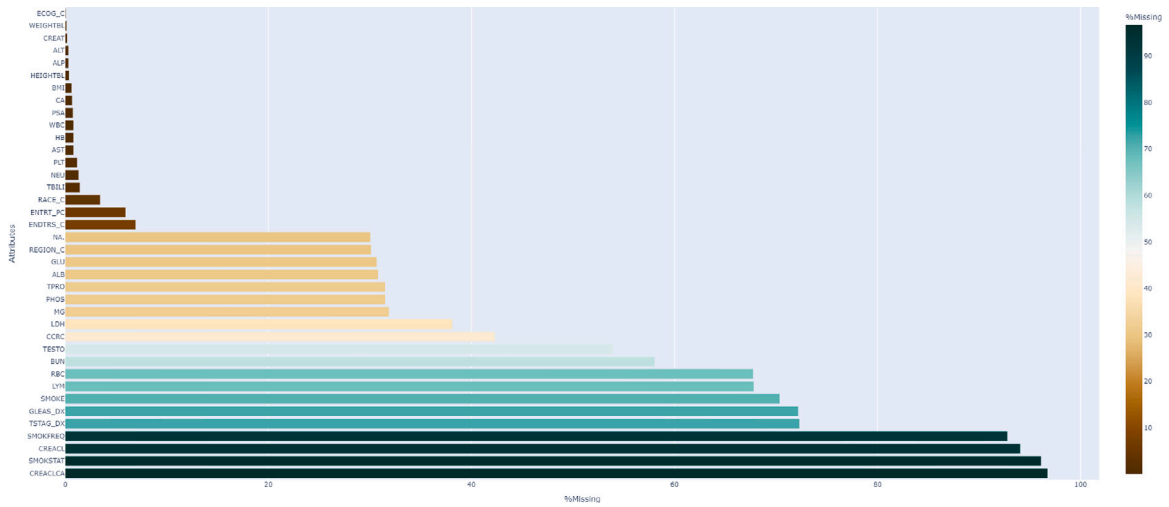

Fig. 1. Rate of missing values per attributes.

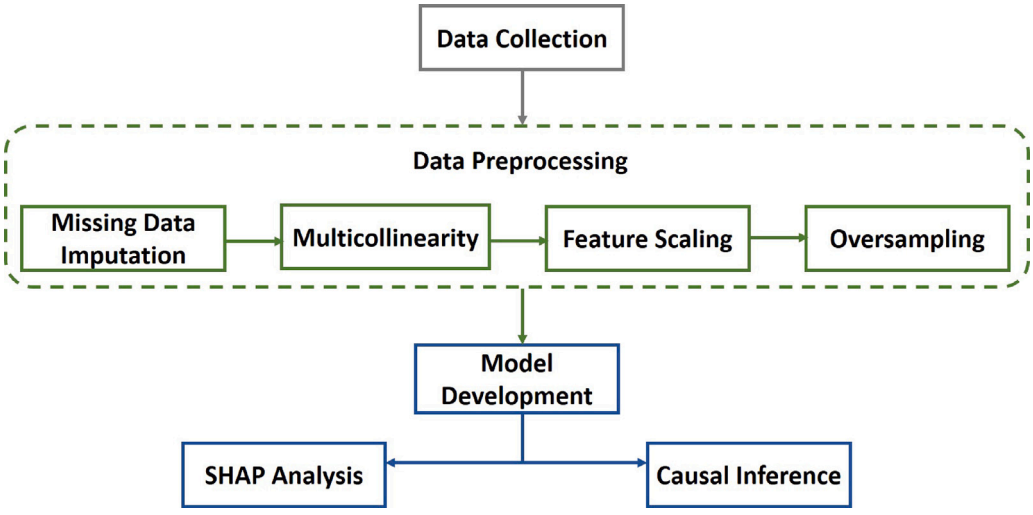

Fig. 2. A flow diagram depicting the workflow.

3.1.3. Feature scaling

Due to the irregularity of the range of values in each column, a large ranged feature can dominate another feature with small range, thereby affecting the contribution of each column to model fitting. To address this, the features were scaled using the Min-Max feature scaling method. The min–max method, also known as unity-based normalization helps to keep the range of values of the data between the points [0,1]. The min–max feature scaling method for each column is calculated as shown in Eq. (1).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where: X is a feature vector, X_{max} is the maximum value in the column, and X_{min} is the minimum value in the column.

3.1.4. Oversampling

An imbalance in data sample labels is a problem, especially with analyses involving health data. It may result in misleading performance of the model, especially when accuracy is needed. To build a robust model, it is important to train the model based on balanced data. Oversampling is a technique which can be used to increase the number of minority samples. It is better than undersampling since undersampling can cause important information to be discarded due to the reduction of the majority class. In this study, we use Adaptive Synthetic algorithm (ADASYN) [33], as explained in Appendix B. This algorithm helps to overcome overfitting as it does not simply duplicate the minority class, but generates samples that are harder to learn.

3.2. Machine learning algorithms

We used eight different machine learning algorithms for analysis. We focused on the categorical boosting (CB), (also known as Catboost algorithm) for the SHAP analysis, and random forest (RF) was used in the causal inference analysis (CausalForest is an extension of random forest). Apart from the aforementioned, the other algorithms used for comparison are support vector machine (SVM), KNN, gaussian naïve bayes (GNB), gradient boosting machine (GBM), logistic regression (LR), and gaussian process classifier (GPC). Each of these algorithms possesses hyperparameters which are tuned during training with cross validation.

3.3. SHapley Additive exPlanations (SHAP)

Although sophisticated models are being built for prediction and modeling using high performing machine learning algorithms, getting insights into the operation of the models is just as important for interpretation. Shapley Additive exPlanations (SHAP) value (or Shapely

Value) is a game theory-based framework, which is theoretically well-founded [34]. It is a framework which uses the concept of Explainable AI for interpreting the predictions of machine learning algorithms. SHAP represents the contribution each predictor has in the machine learning model. Its versatility cuts across both local and global interpretation. It can be used to get insight into a single instance prediction, or the prediction derived from the entire data. SHAP differs from the usual permutation feature importance because it is based on the magnitude of the feature attributions whereas permutation feature importance is based on the decrease in model performance. Hence, the sum of the SHAP values for each prediction is the difference between the prediction and the baseline values which makes it consistent. In addition, SHAP has a fast implementation for tree-based models, which makes it possible to compute all the SHAP values required for global model interpretations. The process for computing the SHAP values is explained in Appendix C

3.4. Causal forest

Causal forest is a tree-based machine learning algorithm which is used to estimate causal effects [35]. It is derived from the random forest algorithm [36]. It uses the decision trees structure for the estimation of the conditional average treatment effect (CATE) of a treatment attribute on an outcome attribute. CATE is the difference between expected outcome under treatment and expected outcome under control, conditional on a set of covariates. It is defined as:

$$CATE(x) = E[Y(1) - Y(0)|X = x], \tag{2}$$

where $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control respectively, and X is the vector of covariates.

To estimate the CATE, random forest is trained on the treated and control observations (these are labeled using an indicator attribute) to predict the outcome attribute. The prediction for observation i is given by:

$$F(X_i) = \frac{1}{N_t}\sum_j Y_j I(T_j = 1) + \frac{1}{N_c}\sum_j Y_j I(T_j = 0), \tag{3}$$

where N_t and N_c are the number of treated and control observations, respectively, and $I(T_j = 1)$ and $I(T_j = 0)$ are indicator functions that take the value 1 if observation j is treated or control, respectively. Using the same data, a new forest is created with a randomly permuted treatment attribute. The new forest is trained to predict the outcome attribute using the permuted treatment attribute and the covariates. The prediction of the second forest is obtained thus:

$$F_{perm}(X_i) = \frac{1}{N}\sum_j Y_j I(T_j = T_i) + \frac{1}{(N-1)}\sum_j Y_j I(T_j ! = T_i), \tag{4}$$

where T_i and T_j are the treatment attributes for observation i and j, respectively and N is the total number of observations. Subsequently, the estimate of the CATE for observation i is:

$$CATE_{\hat{X}_i} = F(X_i) - F_{perm}(X_i) \tag{5}$$

3.5. Evaluation metrics

Evaluation of the method's performance is based on the area under the receiver operating curve (AUROC), and the area under the precision recall curve (AUPRC). These evaluation measures take into consideration the sensitivity and specificity of the models and incorporate them into the performance evaluation. The performance of the models under different thresholds is considered in the identification of the best possible method based on different circumstances. To generate the area under the receiver operating curve, the True Positive Rate (TPR) is plotted against the False Positive Rate (FPR), while for the area under precision recall curve, precision is plotted against recall (same as TPR). The TPR, FPR, and Precision are generated based on Eqs. (6)–(8).

$$TPR/Recall = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)} \tag{6}$$

$$FPR = \frac{FalsePositive(FP)}{TrueNegative(TN) + FalsePositive(FP)} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Table 1
Result of algorithms, with "DISCONT" as target attribute.

Algorithms	AUROC	AUPRC
RF	0.9953	0.9804
SVM	0.8918	0.5723
KNN	0.5852	0.2747
GNB	0.7454	0.4996
GBC	0.9970	0.9710
LR	0.9613	0.8533
GPC	0.9595	0.8127
CB	0.9987	0.9917

4. Analysis and result

For all analyses, we carry out the multicollinearity test to remove the correlated attributes iteratively until the VIF is lower than 5. Further preprocessing such as normalization and oversampling are applied after splitting the dataset. Eight different algorithms were used for analysis. These are random forest (RF), gradient boosting classifier (GBC), catboost (CB), support vector machine (SVM), KNN, gaussian process classifier (GPC), logistic regression (LR), and gaussian naïve bayes (GNB). The data was split for training and testing on a 70:30 ratio. Subsequently, a stratified 5-fold cross validation was applied with grid search for the tuning of the models with the optimal parameters. The best fit estimators were eventually applied on the test data for the evaluation of the model. Data analysis was conducted progressively, following a systematic sequence of steps that ensured the robustness and quality of the data used. The progression of these steps is based on the intuitive understanding of the attributes and the practicality of their respective contributions to model relevance (not just performance).

4.1. Multiple imputation and models' performance

Having analyzed the data with the eight algorithms mentioned earlier, we make comparison between three tree-based algorithms and other algorithms. In the first series of analysis progression, the entire dataset was analyzed, with treatment discontinuation "DISCONT" as the target attribute. After removing the multicollinear attributes iteratively, the dataset had 119 attributes. The result of the analysis is shown in Table 1

It is worth noting the performance of the algorithms using the multiple imputation technique for missing data. Previously, in [23,37,38], where single imputation method had been used to resolve the missing data issue, and discontinuation used as the target attribute, performance of the algorithms were significantly lower than the performances recorded in Table 1. We believe that the method of extrapolating and generating values from neighboring samples over a series of iterations, before combination and analyses allows the models to explore that latent data space more thoroughly, compared to traditional and single imputation methods. This approach does a better job of handling missing values, which is quite prevalent in health-related datasets. Moreover, based on the importance associated with the analysis of health data, single imputation which might not necessarily provide insight to the data should be avoided. The result obtained also justifies the advocacy for the greater adoption of multiple imputation. With this,

Table 2
Result of algorithms, with "DISCONT" as target attribute, but excluding death related attributes.

Algorithms	AUROC	AUPRC
RF	0.7129	0.3839
SVM	0.6193	0.2006
KNN	0.5330	0.2551
GNB	0.5378	0.4814
GBC	0.6902	0.3439
LR	0.6427	0.2317
GPC	0.6167	0.2060
CB	0.7094	0.3881

we have been able to show the superiority of multiple imputation as suggested by [23].

Furthermore, we notice that the tree-based classifiers had superior performance compared to other classifiers. This may be attributed to the ability of tree-based algorithms to effectively handle categorical data. However, when we carried out a SHAP analysis, we discovered that two attributes were influencing performance more than other attributes. These attributes are "ENTRT_PC" and "ENDTRS_C".

The next analysis in the progression involved the removal of the following attributes – "LKADT_P", "DEATH", "ENTRT_PC" and "ENDTRS_C". We reason that these attributes highly influence the prediction, however in a technically inefficient manner. The corresponding implication/interpretation of these attributes are whether the patient dies, the last known alive date, the reason for discontinuing the treatment, and the period of discontinuation respectively. Essentially, these are attributes that should not be known before recommending treatment for a new patient, otherwise they will constitute a bias that will adversely affect the validity of prediction. As a result, we exclude these attributes in the process of setting up our model for prediction. We also iteratively removed multicollinear attributes and this leaves us with 115 attributes for analysis. The respective performance of the classifiers after these processes are shown in Table 2.

In Table 2, we realize that the performance of the algorithms was significantly lower, compared to the earlier outcome presented in Table 1. These results show how much the attributes influenced the behavior of the algorithms, even though it will not be robust for model deployment. Invariably, a random survey of our data shows that those who discontinue the treatment did so because of adverse effects or possible adverse effects as recorded in "ENDTRS_C". However, new patients are not required to go through such a process at the first commencement of their treatment. Also, new patients are not dead, and having the "DEATH" attribute for model building reduces the robustness of the models, even if performance is higher.

Subsequently, we consider the gold standard technique of combination of related attributes as stated in [38]. Considering the relationship between the attributes "DEATH", "DISCONT", and "ENDTRS_C", we derive a new attribute as the target attribute. In the "DEATH" attribute, YES indicates that the patient died, while NO indicates otherwise. While in the "DISCONT" attribute, YES indicates that the patient had to discontinue treatment, while NO indicates otherwise. For the "ENDTRS_C" attribute, there are four categories – Complete, Progression, Possible Adverse Effect, and Adverse Effect. We categorized Possible Adverse Effect and Adverse Effect as YES, while Complete and Progression were categorized as NO. In line with the gold standard for the creation of a new target attribute, if "DEATH" is YES, and "DISCONT" is YES, and "ENDTRS_C" is YES, "NEW_VAR" is set as YES, else NO. Hence, excluding "DEATH", "ENDTRS_C", "LKADT_P", "ENTRT_PC", and "DISCONT", we carry out a new analysis with "NEW_VAR" as the target attribute. The performance of the methods is shown in Table 3. With the removal of multicollinear attributes, there remained 115 attributes for analysis.

As a result of the gold standard, there is a noticeable increase in the performance of the model. More importantly, the AUPRC of the models are higher, which signifies the ability of the model to predict the positive samples. That is, cases whereby the specific use of the treatment can result in either death, adverse effect, or eventual discontinuation are easily detected by the model. As AUPRC is a trade-off between precision and recall, the importance of using the gold standard for analysis is shown.

Table 3
Result of algorithms, with "NEW_VAR" as target attribute, and exclusion of death related attributes.

Algorithms	AUROC	AUPRC
RF	0.6731	0.8944
SVM	0.6260	0.8672
KNN	0.5123	0.8356
GNB	0.6080	0.8604
GBC	0.6422	0.8831
LR	0.6328	0.8760
GPC	0.5931	0.8469
CB	0.6773	0.8941

Table 4
Result of algorithms, with independent test data.

Algorithms	AUROC	AUPRC
RF	0.6335	0.8182
SVM	0.5811	0.8103
KNN	0.4993	0.7728
GNB	0.5713	0.8008
GBC	0.5995	0.8103
LR	0.6632	0.8403
GPC	0.6049	0.8229
CB	0.6595	0.8424

Although the performance recorded in Table 3 is lower than that of Table 1, we consider the models to be more robust and accurate compared to previous ones in literature and in Tables 1–2. For robustness and usability during model deployment, it is important to use attributes that are not records of the aftermath being prevented. This analysis brings into consideration all the possible aftermath that should be prevented and considers it for model development. In addition, the ability of the models to detect positive cases is especially important in health data analysis, where false positives should be avoided. Overall, the importance of multiple imputation techniques and the appropriate use of attributes for modeling has been shown.

To further show the robustness of the model, despite its performance on a multicohort data, we perform independent testing by making one cohort of the data the test set, and training on two other cohort. As shown in Table 4, the robustness of the model is established, due to its performance in an independent test condition.

4.2. SHapley Additive exPlanations (SHAP)

SHAP provides insight into the heuristic mechanism behind the operation of the ML models. It is an important concept of eXplainable AI (XAI), which enables the understanding of the factors influencing model prediction [39]. In this study, a summary plot is shown for the analyses recorded in Tables 1–3, and an additional beeswarm plot for Table 3, since it is the most robust data with reliable attributes for this study. The SHAP algorithm was modeled with Categorical Boosting algorithm.

Fig. 3 shows "ENTRT_PC" as the attribute which mostly influences the decision of the model for prediction. In fact, its mean SHAP value is higher than the sum of the values for the top three attributes. As an aggregate of influencing attributes for all the instances, this shows that the models largely based their predictions on the value of "ENTRT_PC" and "ENDTRS_C". Probing further into the connotation of the top two attributes shows that, since "ENTRT_PC" shows the duration of a patient taking treatment, lower duration means that the

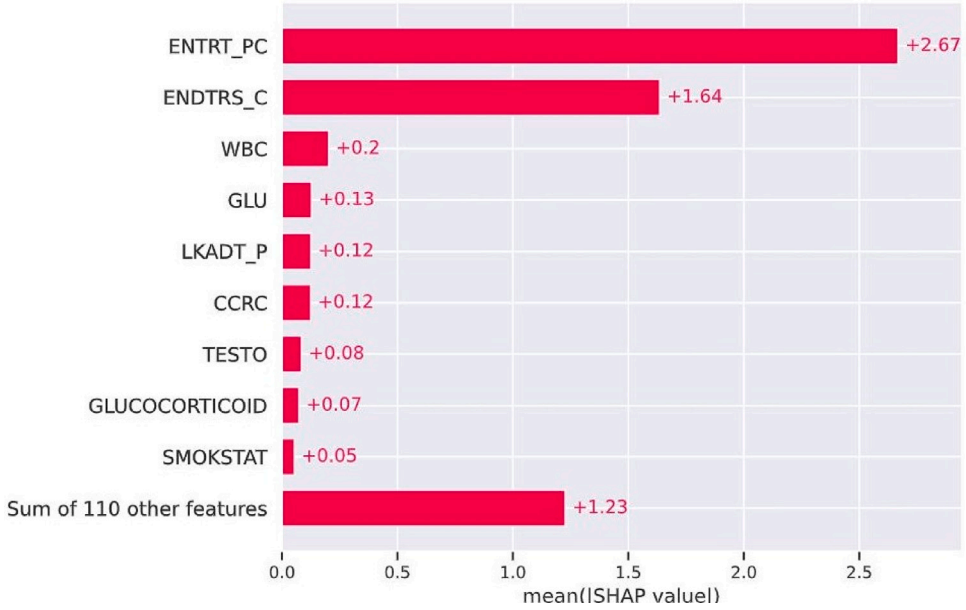

Fig. 3. Mean absolute SHAP value comparison of analysis with "DISCONT" as target attribute.

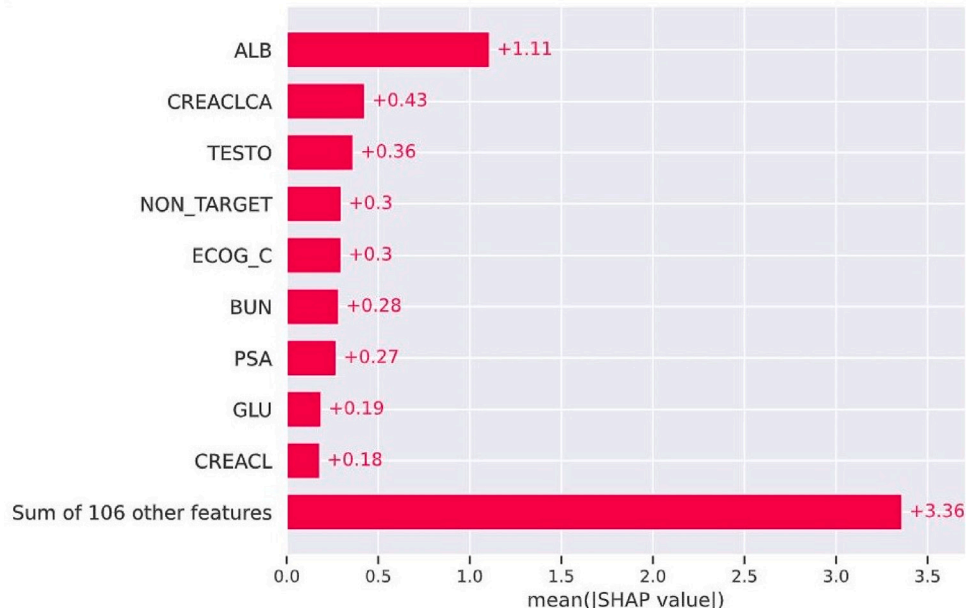

Fig. 4. Mean absolute SHAP value comparison of analysis with "DISCONT" as target attribute but excluding death related attributes.

patient discontinued treatment, while higher numbers tend to show that a patient made progress or completed the treatment. Although the performance of the models is higher as shown in Table 1, this is however not a true reflection of a model to be built for future prognosis. Without the SHAP explanation, it might have been difficult to realize the major impact of these attributes on the prediction analysis, which could result in the development of inefficient, non-robust models.

With the removal of the death-related attributes, Fig. 4 shows that attributes that are available from patient's record before commencement of treatment play an important role in the heuristic process behind the prediction of treatment discontinuation. We also deduce that on the aggregate, the majority of the important attributes with higher mean SHAP values are related to the baseline laboratory value of the patients, compared to demographics, location of metastasis, and their medical history.

Fig. 5 was generated at a point where a new attribute "NEW_VAR" is being used as the target attribute. A noticeable difference in the type of attributes influencing prediction is observed. Now, there is a combination of both clinical lab values, and demographics. The figure reflects the aggregate of the SHAP value of attributes in the gold standard dataset, excluding death related attributes. We notice a fair distribution in the contribution of each attribute to the overall prediction of adverse events (adverse effect, discontinuation, and death) for the patients. According to [40], the decline in testosterone level after a certain age is a factor that determines the incidence of prostate cancer, considering that prostate cancer increases with age exponentially. This is corroborated by Fig. 5 where "TESTO" – testosterone lab value, and "AGEGRP2" – age group are the top contributing attributes to the determination of treatment risk factor. The Gleason score "GLEAS_DX" is used for cancer staging. A high score typically denotes a high-grade cancer which is already a high-risk case and therefore increases the risk

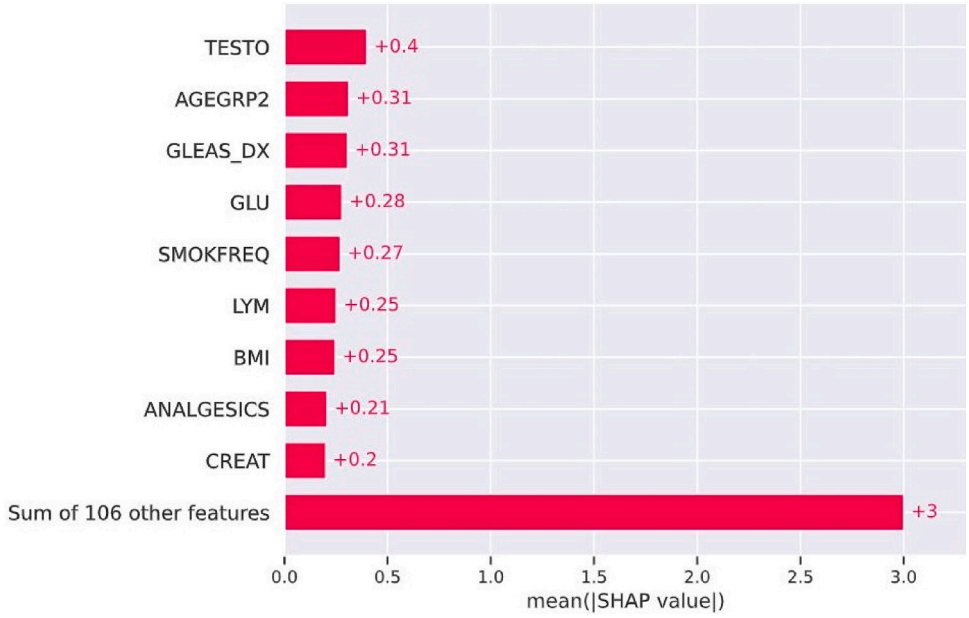

Fig. 5. Mean absolute SHAP value comparison of analysis with "NEW_VAR" as target attribute but excluding death related attributes.

of adverse events. Smoking is known to increase the risk of cancer, and also cause damage to the body. Considering that there is a attribute for smoking, smoking frequency "SMOKFREQ" already indicates that the patient smokes and contributes to the risk level. "LYM" - Neutrophil-to-lymphocyte ratio and platelet-to-lymphocyte ratio are indices used in the diagnosis of prostate cancer and associated with lower survival [41, 42]. It is therefore a biomarker to watch out for in the determination of risk associated with treatment. "BMI" - Having a healthy body mass index (BMI) as an adult is important, as an elevated BMI (tending towards obesity) is associated with prostate cancer risk in men [43], and a higher risk of death from the disease [44]. "ANALGESICS" – Analgesics such as opioids are usually administered to patients who undergo radical prostatectomy, and it is capable of favoring tumor growth by inducing immune suppression and stimulating the proliferation of metastatic prostate cancer [45]. "CREAT" – Creatinine level of patients should be controlled to douse the risk associated with prostate cancer and treatment, as it is known to be positively related to the risk of prostate cancer [46], especially in smokers [47]. Many other factors contribute to the risk of metastatic prostate cancer, and the adverse events associated with treatment.

In Fig. 6, using the beeswarm plot, we can directly see the impact of the value of each attribute on the model output. The plot is similar to that in Fig. 5 as it not only relates to the importance of the contributing attributes and their impacts, but also goes further to expatiate on the impact of the feature value on prediction. Each cluster in the figure is a clump of all the data points and how each independent attribute affected the target attribute. The density of the data points is due to the high number of instances and attributes. The points on the negative side of the x-axis shows a negative impact on prediction, while the points on the right side of the x-axis shows a positive impact on prediction. In addition, points that are farther away from the 0.0 SHAP value are more impactful for prediction compared to those that are close to it. Therefore, for "TESTO", high feature values tend to have a positive impact on adverse events (death, discontinuation, and adverse effect). In addition, "AGEGRP2" shows that although lower age groups still have positive effect on adverse events, the higher age groups are more affected. "SMOKFREQ" – Smoking frequency, and "ANALGESICS" especially denote that people who smoke more or have received analgesics previously, positively contributed to a case of adverse event. A quick glance at Fig. 6 also shows that most of the attributes tend to impact discontinuation and the risk of death. The disadvantage of beeswarm is the clustering of data points which obscures the fine details of the interaction between attributes. As a result, the waterfall model can be used for local expainability as shown in Fig. 7.

In Fig. 7, we use the waterfall plot to show how the attributes affected the prediction of the model for one particular instance. We see that the ranking of the attributes is not necessarily the same as that of Fig. 5, because this is peculiar to one instance. However, in Figs. 5 and 6, we see the aggregation of all the instances used for analysis. The base value on the x-axis, which is the average predicted adverse event across all the samples in the dataset, is $E[f(X)] = -0.173$, while $f(x) = 3.156$ is the predicted adverse event for this sample. The values in the colored arrows are the SHAP values for each of the attributes. Therefore, ANALGESICS is the attribute with most positive contribution to adverse event by 0.48. The same trend follows for other attributes, except the value of "CREAT" decreases the chance of adverse event by 0.2, for this sample. There is a unique waterfall plot for each instance in the dataset, and the aggregate of these plots can generate Figs. 5 and 6.

4.3. Causal investigation

Causal inference is the process of estimating causal effect from an observation. The main finding of causal forests is that for a given observation, the difference between the predictions of the two forests yields an estimate of the CATE. This makes intuitive sense since the first forest discovers the connection between the covariates and the result, whereas the second forest discovers the connection between the covariates and the treatment attribute. We may quantify the causal effect of the treatment attribute on the outcome attribute, conditional on the covariates, by comparing the predictions of the two forests.

Based on a statistical test to determine the significance of the difference between the two forecasts, the causal forest provides an estimate of the CATE. Typically, the statistical test is based on the presumption that the mistakes in the two forests are independent and have identical distributions (i.i.d.). In fact, this assumption is frequently true, but it may not hold true if there is a significant connection between the treatment attribute and the covariates.

For the causal investigation and the analysis of heterogeneity of treatment effect, we implemented the CausalForest using the grf package in R. We loaded the data with the "NEW_VAR" as target, and excluded attributes "DISCONT", "DEATH", "LKADT_P", "ENTRT_PC"

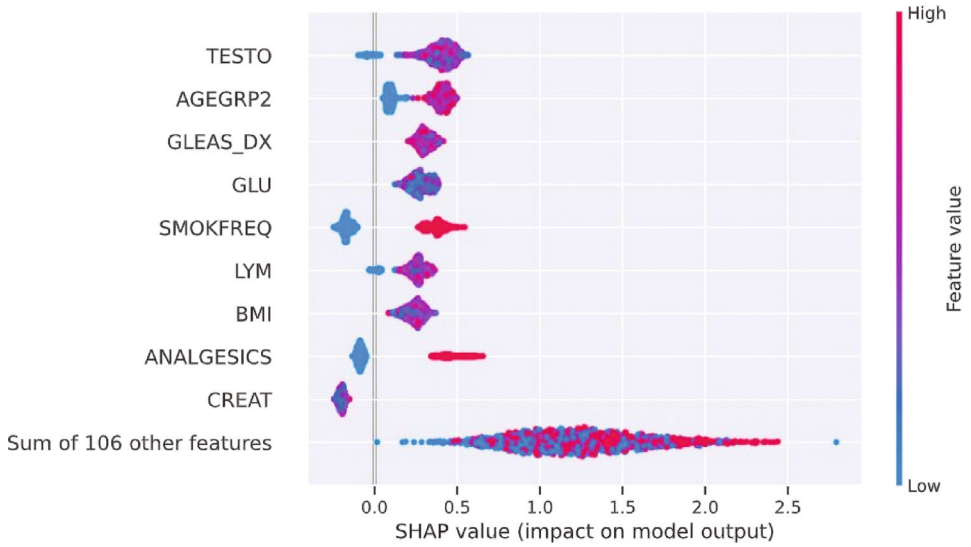

Fig. 6. SHAP beeswarm plot of analysis with "NEW_VAR" as target attribute but excluding death related attributes.

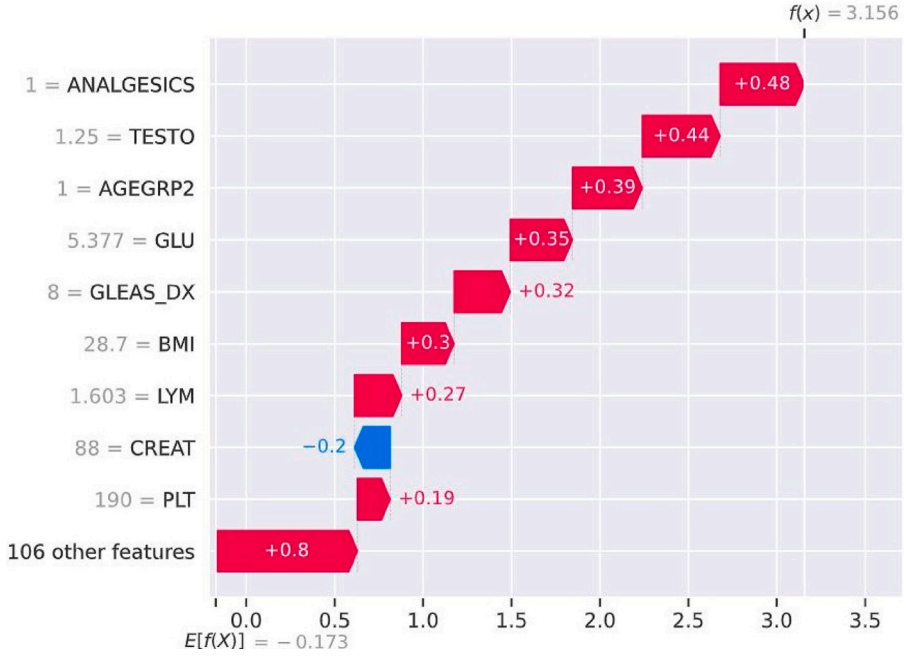

Fig. 7. SHAP waterfall plot of analysis with "NEW_VAR" as target attribute but excluding death related attributes for the local explainability of one patient.

and "ENDTRS_C". On a 70:30 ratio, the data was split into training and test data for the analysis. To train the causal forest with the train data, we set treatment as "SMOKE", since it is a binary attribute which is not influenced by data observer and can be used as intervention. That is, people who smoke as case group, and people who do not smoke as the control group. The mean of the smoke attribute was set as the propensity score, since there is no observed propensity score. The covariates are set as other attributes.

Since CausalForest is based on random forest, majority of its parameters are similar to that of random forest. Firstly, the importance of the attributes was generated, and the attributes greater than the mean of the overall importance was selected. Based on this criterion, 32 attributes were selected for further analysis. Some of the 32 attributes were discovered to be part of the attributes that were also ranked high according to their SHAP values in Fig. 5. These include "TESTO" - Testoterone, "GLEAS_DX", "GLU" - Glucose, "LYM" - Lymphocytes, "BMI", and "ANALGESICS".

To obtain the estimate of the conditional average treatment effect (CATE), the model trained on the training data was used to predict the test data. The summary of the estimate of CATE for the attributes shows the Minimum, 1st Quartile, Median, Mean, 3rd Quartile, and Maximum values as −0.10867, −0.04586, −0.03342, −0.03455, −0.02281, and 0.01620, respectively.

Thereafter, the causal forest was ran on the test data with only the 30 important attributes to obtain the average treatment effect (ATE) given as -0.02872 ± 0.04. The attributes with the best linear projection of the CATE were estimated with confidence intervals that are cluster and heteroskedasticity robust (HC3). We found out that out of the 32 selected attributes, three attributes showed significance with p-value ($p < 0.05$). They are "ANALGESICS", "TESTO" – Testosterone and "CA" – Calcium.

Therefore, we can infer the impact that the previous use of analgesics has on patients who smoke to a reasonable level of confidence. A study by [48] shows that in as much as smoking is associated with

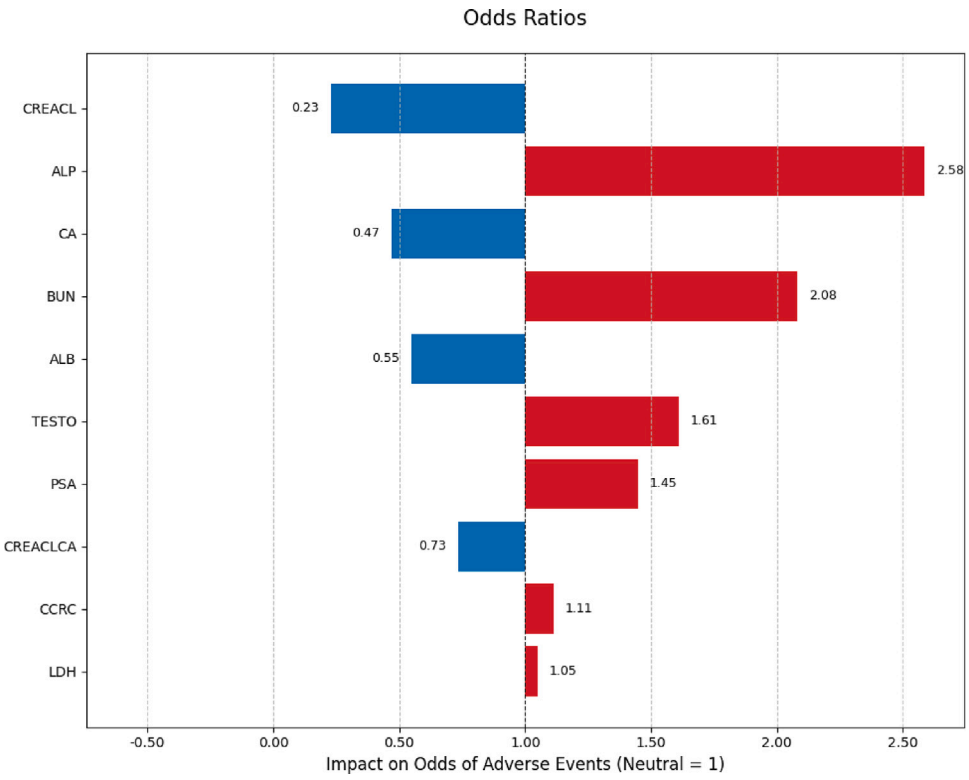

Fig. 8. Tornado diagram showing odds ratio of important features.

various cancers, smokers are at greater risk of chronic postoperative pain than non-smokers, which may result in an inappropriate higher dosage adjustment, and long-term use of analgesics such as opioids. In addition, due to the connection between smoking, pain perception, and the use of opioids (analgesics), smokers are more liable to develop postoperative opioid (analgesics) dependency compared to non-smokers [49,50]. This indicates that healthcare providers need to be wary of analgesic administration to smoking prostate cancer patients, to prevent the triggering of adverse events.

To sum it up, causal forests estimate the CATE of a treatment attribute on an outcome attribute using the decision tree structure. Causal forests can offer a statistically reliable estimate of the CATE by comparing the predictions of two forests trained on the same data with the treatment attribute permuted.

We further carried out a sensitivity analysis to understand how robust the model's conclusions are to changes in assumptions, parameters, or input variables. Using the top 10 feature, we determine the odds ratio of each feature, and plot a corresponding Tornado diagram as shown in Fig. 8. An Odds Ratio quantifies the association between the presence of a factor and the odds of an outcome, when other factors are kept constant. The Neutral Point (OR = 1.0) means the feature has no effect on the odds of the outcome. The Risk Factors (OR > 1.0) indicate features where an increase in that feature is associated with an increased odds of the outcome. The further to the right, the stronger the positive association (higher risk). The Protective Factors (OR < 1.0) indicate features where an increase in that feature is associated with a decreased odds of the outcome. The further to the left, the stronger the negative association (more protective). The length of each bar indicates the magnitude of the effect, with the most influential factors at the top, giving it the tornado shape.

A slight increase in LDH (Lactate Dehydrogenase) is associated with a very slightly increased odds of the adverse event, indicating minimal impact. An increase in CCRC (Calculated Creatinine Clearance) is associated with an 11% increase in the odds of adverse event. An increase in CREACLCA (Cockcroft-Gault Creatinine Clearance) is associated with a 27% decrease in the odds of the adverse event, suggesting it is a protective factor. An increase in PSA (Prostate-Specific Antigen) is associated with a 45% increase in the odds of the adverse event, making it a risk factor. An increase in TESTO (Testosterone) is associated with a 61% increase in the odds of the adverse event, indicating a stronger risk factor. An increase in ALB (Albumin) is associated with a 45% decrease in the odds of the adverse event, making it a significant protective factor. An increase in BUN (Blood Urea Nitrogen) is associated with a 108% increase in the odds of the adverse event, highlighting it as a strong risk factor. An increase in CA (Calcium) is associated with a 53% decrease in the odds of the adverse event, making it a strong protective factor. An increase in ALP (Alkaline Phosphatase) is associated with a 158% increase in the odds of the adverse event, indicating it is the strongest risk factor. An increase in CREACL (Creatinine Clearance) is associated with a 77% decrease in the odds of the adverse event, making it the strongest protective factor.

5. Conclusion

Considering the mortality rate associated with cancer, an in-depth analysis is required to quantify and determine the factors that predispose patients to mortality, especially those with metastasis. In this study, we have performed predictive analysis with an emphasis on tree-based machine learning models. We showed the importance of multiple imputation for dealing with missing data compared to single imputation as used in previous studies. In addition, we conducted a series of analyses to effectively understand the data and generate models that are robust and pertinent to patient diagnosis, removing death-related attributes since patients are expected to be alive to determine the optimal treatment for them.

In addition, we explore the use of eXplainable AI techniques to give an exploratory detail of the attributes that influence decision-making for each patient, and on an aggregate level. This mitigates the obscurity imposed by the 'black-box' nature of ML algorithms, making users 'less blind' to the inner workings of these algorithms. Essentially,

researchers, patients, and medical practitioners have a window of insight into the decision heuristics of these models and make decisions based on the perspective they give. Furthermore, we performed a causal inference analysis of data based on the smoking status of patients and discovered that testosterone level, use of analgesics, and calcium level of patients are attributes that significantly influence adverse events in the treatment of patients who have a smoking history.

This study is subject to certain limitations. First, the use of data from multi cohorts, while advantageous for enhancing the external validity and robustness of our findings, introduces potential heterogeneity and data bias due to varying collection protocols and population characteristics across different sites. Second, as is common in large-scale health analyses, the dataset contained missing values; however, potential bias was mitigated by employing multiple imputation techniques. Despite these rigorous methodological adjustments, interpretations should be made with the understanding that residual confounding or unmeasured variations inherent to multicohort designs may still exist.

The integration of interpretable AI into mCPRC treatment holds transformative potential for bridging the gap between computational precision and decision-making. By leveraging causal inference and SHAP analysis, this study moves beyond "black box" predictions, offering clinicians transparent, actionable insights into risk factors that drive disease progression, and treatment-related adverse events. This interpretability is critical for fostering trust in AI-assisted decisions, particularly when recommending intervention. Furthermore, the model's robustness based on the analysis of the multicohort data suggests that the algorithm can maintain high performance across diverse patient populations and clinical settings. Ultimately, implementing these validated, explainable models in real-world workflows could reduce overtreatment of disease, optimize resource allocation, and help in strategic personalized treatment recommendations.

In subsequent studies, causal analysis can be performed using another attribute to differentiate the control and treatment groups. We have limited this work to the smoking status of patients, but beyond that, treatment and control can be based on several factors. In addition, due to the computational cost or model-size constraints for model deployment in resource limited clinical environments, high-accuracy quantization methods such as IR-QLoRA (Information Retention-Quantized LoRA) [51] can be utilized to ensure ensures that critical clinical performance is retained despite compression. This study provides not just results for machine learning algorithms, but also reasons and insights which can be applicable in the treatment of metastatic cancer patients, for quality health.

CRediT authorship contribution statement

Olutomilayo Olayemi Petinrin: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Faisal Saeed:** Writing – original draft, Funding acquisition. **Hao Xue:** Methodology, Formal analysis. **Sumanta Basu:** Supervision, Methodology. **Shadi Basurra:** Funding acquisition, Formal analysis. **Zhe Liu:** Writing – original draft. **Muhammad Toseef:** Writing – original draft. **Ibukun Omotayo Muyide:** Formal analysis. **Ka-Chun Wong:** Supervision, Funding acquisition.

Funding statement

This research is funded by Data Analytics and Artificial Intelligence (DAAI) research group, School of Computing and Digital Technology, Birmingham City University, UK. This research was substantially sponsored by the research projects (Grant No. 32170654 and Grant No. 32000464) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong. This project was substantially funded by the Strategic Inter-disciplinary Research Grant of City University of Hong Kong (Project No. 2021SIRG036). The work described in this paper was partially supported by the grants from City University of Hong Kong, Hong Kong (CityU 9667265).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Variance inflation factor

Variance Inflation Factor (VIF) is a common tool which is used to detect multicollinearity between independent attributes [52]. It measures the ratio of variance of the coefficients of a model's features based on the inclusion of a particular feature. VIF for each attribute is determined using Eq. (A.1). From Eq. (A.1), R_i^2 is the unadjusted coefficient of determination with which the ith independent attribute is regressed on other attributes.

$$VIF_i = \frac{1}{1 - R_i^2} \tag{A.1}$$

When R_i^2 is 0, it indicates that the variance of other independent attributes cannot be determined from the ith independent attribute. Conversely, when VIF is 1, it indicates that the ith independent attribute is not correlated to other independent attributes, and multicollinearity does not exist. Typically, any attribute with VIF higher than 5 or 10 is contributing to multicollinearity [53]. To tackle this problem, the attribute with high VIF is removed from the data. It is important to iteratively determine the VIF of the attributes and remove one at a time, instead of removing all attributes with high VIF at once. The removal of the attribute has less impact on the coefficient of determination since the information provided by the removed attribute is deemed redundant.

Appendix B. Adaptive Synthetic Algorithm (ADASYN)

To apply Adaptive Synthetic algorithm, the ratio of the minority to majority samples is calculated using Eq. (B.1):

$$Ratio = \frac{n_{min}}{n_{maj}}, \tag{B.1}$$

where n_{min} and n_{maj} are the number of the minority and majority class samples respectively. If the ratio is below a certain threshold, the algorithm is initialized. Thereafter, the total number of synthetic minority class samples to be generated is calculated based on Eq. (B.2)

$$T = (n_{maj} - n_{min})\beta, \tag{B.2}$$

where T represents the total number of synthetic minority samples to be generated, β indicates the desired ratio of minority to majority class samples after oversampling. If $\beta = 1$, a perfectly balanced dataset is expected.

Next, the KNN of each minority sample is located, and the r_i value is calculated using Eq. (B.3). The r_i value shows the dominance of the majority class in each neighborhood. Neighborhoods with higher r_i are more difficult to learn since they contain more majority class samples. At this point, each minority sample is associated with a different neighborhood.

$$r_i = \frac{number\ of\ majority\ samples\ in\ a\ neighborhood}{k} \tag{B.3}$$

Thereafter, the r_i values are normalized as shown in Eq. (B.4) such that the sum is equal to 1.

$$\hat{r}_i = \frac{r_i}{\sum r_i}, \text{and} \sum \hat{r}_i = 1 \tag{B.4}$$

The number of synthetic samples to be generated per neighborhood is $T_i = T_{\hat{r}_i}$. Since r_i is higher in majority class dominated neighborhoods, those neighborhood will have more synthetic minority class samples generated, which describes the adaptive nature of ADASYN due to the generation of samples in neighborhoods that are harder to learn. Now, T_i data for each neighborhood can be generated. First, the minority

sample x_i for the neighborhood is taken, then another minority sample x_j within the same neighborhood is selected. Based on this, the new synthetic sample is calculated using Eq. (B.5). where s_i is the new synthetic sample, λ is a random number between 0 and 1, and x_i, x_j are minority samples within the same neighborhood.

$$s_i = x_i + (x_j - x_i)\lambda \tag{B.5}$$

Appendix C. SHAP

To get the global importance I_j, the absolute SHAP value per feature across data is averaged based on Eq. (C.1).

$$I_j = \frac{1}{n}\sum_{i=1}^{n} |\phi_j^{(i)}| \tag{C.1}$$

To determine the SHAP values for a given prediction of an instance, the weights are determined by the conditional expectations of the model's predictions given the presence or absence of each feature as given in Eq. (C.2). where f_i is the contribution of the ith feature to the prediction, w_i is the weight of the ith feature, determined by the conditional expectations of the model's predictions given the presence or absence of each feature, and $E[f|x_{(-i)}]$ is the expected prediction value, given the absence of the ith feature.

$$SHAP_i = (f_i - E[f|x_{(-i)}]) * w_i \tag{C.2}$$

References

[1] Bacac M, Stamenkovic I. Metastatic cancer cell. Annu Rev Pathol: Mech Dis 2008;3(1):221–47. http://dx.doi.org/10.1146/annurev.pathmechdis.3.121806.151523, URL https://www.annualreviews.org/doi/abs/10.1146/annurev.pathmechdis.3.121806.151523.

[2] Bansal A, Simon MC. Glutathione metabolism in cancer progression and treatment resistance. J Cell Biol 2018;217(7):2291–8.

[3] Steeg PS. Targeting metastasis. Nat Rev Cancer 2016;16(4):201–18.

[4] Siegel RL, Kratzer TB, Giaquinto AN, Sung H, Jemal A. Cancer statistics, 2025. Ca 2025;75(1):10.

[5] Wani SA, Qudrat S, Zubair H, Iqbal Z, Gulzar B, Aziz S, Inayat A, Safi D, Kamran A. Role of osteoclast inhibitors in prostate cancer bone metastasis; a narrative review. J Oncol Pharm Pr 2024;10781552241275943.

[6] Prete AA, Angerilli V, Bergamo F, Vettore V, De Toni C, Intini R, Cerma K, Ricagno G, Cerantola R, Perissinotto E, et al. HER2 expression and genOmic characterization of rESected brain metastases from colorectal cancer: The HEROES study. Br J Cancer 2024;130(8):1316–23.

[7] National Cancer Institute. Metastatic Cancer: When Cancer Spreads. U. S. National Institutes of Health, National Cancer Institute; 2020, URL https://www.cancer.gov/types/metastatic-cancer#research, [Accessed 10 December 2024].

[8] Zhao SG, Bootsma M, Zhou S, Shrestha R, Moreno-Rodriguez T, Lundberg A, Pan C, Arlidge C, Hawley JR, Foye A, et al. Integrated analyses highlight interactions between the three-dimensional genome and DNA, RNA and epigenomic alterations in metastatic prostate cancer. Nature Genet 2024;56(8):1689–700.

[9] Brady L, Kriner M, Coleman I, Morrissey C, Roudier M, True LD, Gulati R, Plymate SR, Zhou Z, Birditt B. Inter-and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. Nat Commun 2021;12(1):1–16.

[10] Iyengar V, Agrawal S, Chiasakul T, Tehranchi K, Mcnichol M, Carney BJ, Leader A, Zwicker JI, Patell R. Comparison of direct oral anticoagulants versus low-molecular-weight heparin in primary and metastatic brain cancers: A meta-analysis and systematic review. J Thromb Haemost 2024;22(2):423–9.

[11] Amato O, Giannopoulou N, Ignatiadis M. Circulating tumor DNA validity and potential uses in metastatic breast cancer. NPJ Breast Cancer 2024;10(1):21.

[12] Hao X, Jiang B, Wu J, Xiang D, Xiong Z, Li C, Li Z, He S, Tu C, Li Z. Nanomaterials for bone metastasis. J Control Release 2024;373:640–51.

[13] Bonni S, Brindley DN, Chamberlain MD, Daneshvar-Baghbadorani N, Freywald A, Hemmings DG, Hombach-Klonisch S, Klonisch T, Raouf A, Shemanko CS, et al. Breast tumor metastasis and its microenvironment: It takes both seed and soil to grow a tumor and target it for treatment. Cancers 2024;16(5):911.

[14] Housini M, Dariya B, Ahmed N, Stevens A, Fiadjoe H, Nagaraju GP, Basha R. Colorectal cancer: Genetic alterations, novel biomarkers, current therapeutic strategies and clinical trials. Gene 2024;892:147857.

[15] Patwekar M, Sehar N, Patwekar F, Medikeri A, Ali S, Aldossri RM, Rehman MU. Novel immune checkpoint targets: A promising therapy for cancer treatments. Int Immunopharmacol 2024;126:111186.

[16] Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018;15(2):81.

[17] Zhang J, Cunningham JJ, Brown JS, Gatenby RA. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. Nat Commun 2017;8(1):1–9.

[18] Seruga B, Ocana A, Tannock IF. Drug resistance in metastatic castration-resistant prostate cancer. Nat Rev Clin Oncol 2011;8(1):12–23.

[19] Li H, Liu R-B, Long C-m, Teng Y, Liu Y. A novel machine learning prediction model for metastasis in breast cancer. Cancer Rep 2024;7(3):e2006.

[20] Chen Q, Chen J, Deng Y, Bi X, Zhao J, Zhou J, Huang Z, Cai J, Xing B, Li Y, et al. Personalized prediction of postoperative complication and survival among colorectal liver metastases patients receiving simultaneous resection using machine learning approaches: A multi-center study. Cancer Lett 2024;593:216967.

[21] Guo Z, Zhang Z, Liu L, Zhao Y, Liu Z, Zhang C, Qi H, Feng J, Yang C, Tai W, et al. Machine learning for predicting liver and/or lung metastasis in colorectal cancer: A retrospective study based on the SEER database. Eur J Surg Oncol 2024;50(7):108362.

[22] Cai W-l, Cheng M, Wang Y, Xu P-h, Yang X, Sun Z-w, Yan W-j. Prediction and related genes of cancer distant metastasis based on deep learning. Comput Biol Med 2024;168:107664.

[23] Petinrin OO, Li X, Wong K-C. Particle swarm optimized Gaussian process classifier for treatment discontinuation prediction in multicohort metastatic castration-resistant prostate cancer patients. IEEE J Biomed Health Informatics 2021;26(3):1309–17.

[24] Liu M-W, Zhang X, Wang Y-M, Jiang X, Jiang J-M, Li M, Zhang L. A comparison of machine learning methods for radiomics modeling in prediction of occult lymph node metastasis in clinical stage IA lung adenocarcinoma patients. J Thorac Dis 2024;16(3):1765.

[25] Mahmoodifar S, Pangal DJ, Neman J, Zada G, Mason J, Salhia B, Kaisman-Elbaz T, Peker S, Samanci Y, Hamel A, et al. Comparative analysis of the spatial distribution of brain metastases across several primary cancers using machine learning and deep learning models. J Neurooncol 2024;167(3):501–8.

[26] Wang H, Zhou Z, Li Y, Chen Z, Lu P, Wang W, Liu W, Yu L. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 F-FDG PET/CT images. EJNMMI Res 2017;7(1):1–11.

[27] Project Data Sphere. Metastatic Castration Resistant Prostate Cancer Data, project data sphere. 2021, URL https://data.projectdatasphere.org/projectdatasphere/html/home, [Accessed 02 January 2023].

[28] Loo RTJ, Pavelka L, Mangone G, Khoury F, Vidailhet M, Corvol J-C, Glaab E. Multi-cohort machine learning identifies predictors of cognitive impairment in Parkinson's disease. NPJ Digit Med 2025;8(1):482.

[29] Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol 2018;18(1):1–16.

[30] Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiology 2017;9:157.

[31] Woods AD, Gerasimova D, Van Dusen B, Nissen J, Bainter S, Uzdavines A, Davis-Kean PE, Halvorson M, King KM, Logan JA, et al. Best practices for addressing missing data through multiple imputation. Infant Child Dev 2024;33(1):e2407.

[32] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? Int J Methods Psychiatr Res 2011;20(1):40–9.

[33] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE; 2008, p. 1322–8.

[34] Winter E. The shapley value. Handb Game Theory Econ Appl 2002;3:2025–54.

[35] Athey S, Wager S. Estimating treatment effects with causal forests: An application. Obs Stud 2019;5(2):37–51.

[36] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[37] Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, Greiner R, Fuchs C, Gofer E, Kumar L, Wolfinger RD, et al. A DREAM challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. JCO Clin Cancer Informatics 2017;1:1–15.

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

[39] Petinrin OO, Saeed F, Toseef M, Liu Z, Basurra S, Muyide IO, Li X, Lin Q, Wong K-C. Machine learning in metastatic cancer research: Potentials, possibilities, and prospects. Comput Struct Biotechnol J 2023.

[40] Xu X, Chen X, Hu H, Dailey AB, Taylor BD. Current opinion on the role of testosterone in the development of prostate cancer: A dynamic model. BMC Cancer 2015;15:1–8.

[41] Cao J, Zhu X, Zhao X, Li X-F, Xu R. Neutrophil-to-lymphocyte ratio predicts PSA response and prognosis in prostate cancer: A systematic review and meta-analysis. PLoS One 2016;11(7):e0158770.

[42] Xu Z, Zhang J, Zhong Y, Mai Y, Huang D, Wei W, Huang J, Zhao P, Lin F, Jin J. Predictive value of the monocyte-to-lymphocyte ratio in the diagnosis of prostate cancer. Medicine 2021;100(38).

[43] Lavalette C, Cordina Duverger E, Artaud F, Rébillard X, Lamy P-J, Trétarre B, Cénée S, Menegaux F. Body mass index trajectories and prostate cancer risk: Results from the EPICAP study. Cancer Med 2020;9(17):6421–9.
[44] Tzenios N, Tazanios ME, Chahine M. The impact of body mass index on prostate cancer: An updated systematic review and meta-analysis. Medicine 2022;101(45):e30191.
[45] Lee BM, Singh Ghotra V, Karam JA, Hernandez M, Pratt G, Cata JP. Regional anesthesia/analgesia and the risk of cancer recurrence and mortality after prostatectomy: A meta-analysis. Pain Manag 2015;5(5):387–95.
[46] Gu X, Wu J, Liu X, Hong Y, Wu Y, Tian Y. Role of serum creatinine levels in prognostic risk stratification of prostate cancer patients. Med Sci Monit: Int Med J Exp Clin Res 2022;28:e937100–1.
[47] Weinstein SJ, Mackrain K, Stolzenberg-Solomon RZ, Selhub J, Virtamo J, Albanes D. Serum creatinine and prostate cancer risk in a prospective StudySerum creatinine and prostate cancer. Cancer Epidemiol Biomarkers Prevent 2009;18(10):2643–9.
[48] Kaptur A, Dziedziński D, Łoś D, Nowak A, Janus A. The impact of smoking on general anesthesia: A literature review. Qual Sport 2024;16:52526.
[49] Etcheson JI, Gwam CU, George NE, Walia N, Jerjian C, Han G-r, Virani S, Miller SJ, Delanois RE. Opiate pain medication consumption in cigarette smokers following total hip arthroplasty. Joints 2018;6(03):157–60.
[50] Yoo SH, Kang J, Kim HJ, Lee SW, Hong M, Jung EH, Kim YJ, Yon DK, Kang B. Opioid use and subsequent delirium risk in patients with advanced cancer in palliative care: A multicenter registry study. Sci Rep 2024;14(1):6004.
[51] Qin H, Ma X, Zheng X, Li X, Zhang Y, Liu S, Luo J, Liu X, Magno M. Accurate lora-finetuning quantization of llms via information retention. 2024, arXiv preprint arXiv:2402.05445.
[52] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol. 112, Springer; 2013.
[53] Garcia Garcia C, Salmeron Gomez R, Garcia Perez J. A review of ridge parameter selection: Minimization of the mean squared error vs. mitigation of multicollinearity. Comm Statist Simulation Comput 2024;53(8):3686–98.