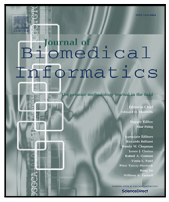
Journal of Biomedical Informatics 141 (2023) 104365

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Explainable discovery of disease biomarkers: The case of ovarian cancer to illustrate the best practice in machine learning and Shapley analysis

Weitong Huang [a,*], Hanna Suominen [a,b], Tommy Liu [a], Gregory Rice [c,d], Carlos Salomon [d,1], Amanda S. Barnard [a,1]

[a] School of Computing, Australian National University, Acton, ACT 2601, Australia
[b] Department of Computing, University of Turku, Turku, Finland
[c] Inoviq Limited, Notting Hill, Australia
[d] Translational Extracellular Vesicles in Obstetrics and Gynae-Oncology Group, Centre for Clinical Diagnostics, University of Queensland Centre for Clinical Research, Royal Brisbane and Women's Hospital, Faculty of Medicine, The University of Queensland, Brisbane, Australia

ARTICLE INFO

Keywords:
Cancer screening
Supervised machine learning
Medical informatics
Evaluation study as topic

ABSTRACT

Objective: Ovarian cancer is a significant health issue with lasting impacts on the community. Despite recent advances in surgical, chemotherapeutic and radiotherapeutic interventions, they have had only marginal impacts due to an inability to identify biomarkers at an early stage. Biomarker discovery is challenging, yet essential for improving drug discovery and clinical care. Machine learning (ML) techniques are invaluable for recognising complex patterns in biomarkers compared to conventional methods, yet they can lack physical insights into diagnosis. eXplainable Artificial Intelligence (XAI) is capable of providing deeper insights into the decision-making of complex ML algorithms increasing their applicability. We aim to introduce best practice for combining ML and XAI techniques for biomarker validation tasks.
Methods: We focused on classification tasks and a game theoretic approach based on Shapley values to build and evaluate models and visualise results. We described the workflow and apply the pipeline in a case study using the CDAS PLCO Ovarian Biomarkers dataset to demonstrate the potential for accuracy and utility.
Results: The case study results demonstrate the efficacy of the ML pipeline, its consistency, and advantages compared to conventional statistical approaches.
Conclusion: The resulting guidelines provide a general framework for practical application of XAI in medical research that can inform clinicians and validate and explain cancer biomarkers.

1. Introduction

There has been little improvement in survival rates of ovarian cancer over the last 20 years. The major contributing factor to the high mortality rate is the lack of clinically useful biomarkers for earlier detection of ovarian cancer, creating an urgent need for non-invasive, specific biomarkers to identify patients at early stages. While no ovarian cancer test has been developed that is suitable for community-based screening, evidence supports the hypothesis that certain epithelial ovarian cancers may be detectable up to two years prior to their clinical presentation [1–3]. Further advances will likely come from analysis of historical data.

Machine learning (ML) [4] is an effective and efficient tool for analysing the vast amount of data of various modalities generated from abundant sources in medical science and health care [5–8]. Medical researchers are exploring ML approaches to analyse data on a large scale to identify previously unknown patterns that could lead to discovering clinically useful disease biomarkers [9–13]. However, with the rapid increase in advanced ML methods, it is not immediately clear which methods should be used and how domain experts should apply them correctly.

Biomarkers have received considerable attention because they show significant potential to improve outcomes in applications including drug discovery and clinical care [14,15]. Cancer biomarker development includes five key phases [16], including Phase (1) Preclinical Exploratory Studies, Phase (2) Clinical Assay Development for Clinical Disease, Phase (3) Retrospective Longitudinal Repository Studies, Phase (4) Prospective Screening Studies, and Phase (5) Clinical Outcome Studies. In the second phase, a clinical biomarker assay is established. This is where ML models may have the most impact by discovering hidden patterns in data that are useful for distinguishing different data

* Corresponding author.
E-mail address: jacob.huang@anu.edu.au (W. Huang).
[1] Senior authorship.

https://doi.org/10.1016/j.jbi.2023.104365
Received 22 December 2022; Received in revised form 24 March 2023; Accepted 10 April 2023
Available online 14 April 2023

labels (the diagnosis) based on the training set ground truth classes (the true presence or absence of the disease) that data instances (patients) belong in a specific task (prediction).

In biomarker discovery, ML methods may provide a more systematic way of making classifications compared to empirical methods, such as the Risk of Ovarian Malignancy Algorithm (ROMA) metric that uses logistic regression [17–21]. Better screening biomarkers are urgently needed to identify more patients at an early stage, exceeding current success rates of only 30% to 45% [22,23]. ML, however, can suffer from the "black-box" problem [24–26], which refers to situations where the results cannot be explained. This includes understanding the structure of the model itslef, and the degree to which domain experts (who are ultimately accountable) can trust the result and act on it accordingly. This problem is exacerbated by the extreme complexity of the more accurate ML methods that makes them harder to interpret or explain, and is highly relevant to both ML and medical domain experts [27–30]. Research into explaining black-box models is referred to as eXplainable AI (XAI), and presents a number of opportunities that can be leveraged by the medical community [25,31].

While it is preferable to predict the risk of presenting with cancer, *a priori* cancer diagnosis is critical in complicated cases such as ovarian cancer where more effective early detection assays are yet to be established. Diagnosing cancer can be posed as a classification problem where the goal is to build a high-quality classifier to predict positive cases using biomarkers as variables. The quality of a classifier can be measured by the sensitivity and specificity (or the area under the curve of receiver operating characteristic, AUC-ROC score). In sensitive medical applications, explainability is often required and the need to understand the model can surpass the need for classification ability. Choosing suitable models and analysis methods is therefore important, since it ultimately balances the outcome explainability and classification ability [32]. In general, less complex models (e.g., logistic regression [33] or decision trees [34]) are easier to explain but have poor classification ability compared to more complex models (e.g., neural networks [35] or random forests [36]). Classification ability and explainability also depends on the data used, but a better understanding of these capabilities can guide method selection. Data containing relatively "simple" patterns may do not require a sophisticated algorithm, and complex models may lead to unexpected overfitting [37]. When this is the case, post-hoc analysis methods such as SHapley Additive exPlanation (SHAP) [38] can improve the understanding of complex models in a straightforward way that could inform the re-selection of a more appropriate ML model.

The aim of this paper is to establish a best practice for selecting and applying ML methods for developing diagnostic assays that make use of tabular data, with a particular emphasis on XAI. As we will show, XAI can be especially valuable to clinicians as it provides straightforward domain knowledge explanations for both biomarker discovery and disease diagnosis. General ML workflows and guidelines are most applicable to the second phase of biomarker development: Clinical Assay Development, as identified by Pepe et al. [16]. Specifically, we focus on model evaluation and analysis using SHAP, a game theoretic XAI approach based on Shapley values, to address accountability issues from the medical, computer science, and even legislative perspectives [39–44].

Our study results can be summarised in two ways. First, we propose a guideline for adopting ML methods in biomarker discovery tasks that takes the need for accountability into consideration when performing the prediction task using XAI. Second, we demonstrate our pipeline on the public CDAS PLCO Ovarian Biomarkers dataset [45] and obtain results with both classification ability and explainability consistent with previous literature. By using Shapley values in biomarker validation tasks, we show how a domain understanding of ML algorithms is straightforward to establish.

2. Methodology

A general ML pipeline includes the following steps: (1) data collection, (2) data preprocessing, (3) model selection, (4) model training, (5) result evaluation, and (6) application [46,47]. While there is no "one best" approach for all problems, we summarise techniques and common pitfalls and propose general steps for researchers to consider when applying ML to cancer biomarker discovery. We also highlight where and how to include XAI.

2.1. Data collection

Data collection consists of data acquisition, data characterisation (often referred to as feature extraction), data cleaning and processing, data labelling and data improvement (often referred to as feature engineering) [48].

To make data collection explainable, it is preferable to actively use domain knowledge. Meta-data, such as the inclusion criteria for a patient cohort, key dates, measurement methods, asset numbers of details of instrumentation and attribute descriptions, inform subsequent decisions, such as choosing feature engineering criteria or base models. For example, if an underlying trend based on the date a sample was taken is discovered, this can be accounted for in a model, or removed in a calibration process.

The PLCO Ovarian Biomarkers dataset was well established via clinical trials, where 113 cases and 894 non-cases exist after removing missing values. With abundant literature discussing the biomarkers recorded in this dataset [17,49–52], there is sufficient domain knowledge. The dataset has rigorous inclusion criteria, variable descriptions, protocols and many domain specific metadata that provided further modelling processes with a transparent background for examination.

Ethical approval (Protocol 2022/261) was obtained from the Australian National University Human Research Ethics Committee to use the dataset for the purposes of this study. No new data were collected.

2.2. Data pre-processing

While actions such as dealing with missing data, data integration and data standardisation fall under the data pre-processing category, dimensionality reduction techniques, such as feature selection, have been given greater attention because they closely relate to the model training and evaluation, and they directly affect explainability since variables could be removed from a dataset. Variable removal often implies lower correspondence with the classification goal, but can also have significant implications on how the ML model uses the remaining variables.

Data preprocessing should only be performed on the training set, with the test set left out. Missing values are either interpolated or, as in this study due to their random distribution, removed. Variables in this study were also normalised to have a consistent range. Implications of all variables recorded in the dataset have been discussed in the domain literature [20,50,53–55]. To make the most of the information in this case study, two variables were removed as they were highly correlated (>90%) with other variables in the dataset (removing *panelc_igfbpii_nml_lp_log* and *panelc_slpi_log* to retain *panelc_igfbpii_nml_lp* and *panelc_slpi*). This step reduced redundancy while retaining the most useful information.

2.2.1. Dimensionality reduction

Dimensionality reduction is used to balance classification ability and computational resource requirements. High dimensionality risks increasing the training time and overfitting the result, while low dimensionality risks impairing the classification ability due to a lack of information. Methods for dimensionality reduction includes feature extraction and feature selection [31,32]. For medical applications the choice of dimensionality reduction method is important since feature

extraction methods, such as Principle Component Analysis (PCA) [56], are generally not explainable, while feature selection methods retain the potential for explainability. Choosing the right variables can enhance the explainability of a model by providing a clearer relationship between variables and classification goals, while also reducing the cost of data collection and model training.

Feature selection can be guided by domain knowledge. It is very likely that biomarkers have non-linear relationships with the classification goal and more complex models are required for deeper analysis. For example, if a uni-variate feature selection algorithm [57] is adopted without considering domain knowledge implications, more nuanced non-linear variables may be automatically eliminated. Recovering and retaining those variables would increase explainability, even if the model's classification ability was unaffected.

At this point it is important to note that any feature selection algorithm should be determined based on the training dataset only [58]. The training and test datasets should contain consistent variable types although their values might have different distributions. A good feature selection criterion should be universal on the population and this can only be achieved by ensuring test set is not exposed to the process.

In our case study, we apply a random forest classifier to select the top 20 variables with highest relative feature importance (FI) [59] from the dataset and use only those for the modelling process (Table B.4). A random forest classifier provides inherent feature importance profiles from its training result. Compared to other models, such as logistic regression or decision tree, that also generate such profiles, a random forest has the advantage of involving randomness in the process, which makes the result more general. Choosing 20 variables balances the need for reducing the dimensionality and considering potential contributions of statistically less significant variables. Other choices could be made depending on computational resource limitations.

2.3. Model training and evaluation

After the training and test data split, a further split of the whole training population into training data and validation data specifically for each training iteration is essential. The term "validation", as used by many medical researchers as the final performance evaluation is similar to "testing" in ML (e.g., see differences in [60] and [61]). In ML, however, it refers a process where intermediate models are "validated" on part of the training dataset hidden for that optimisation iteration. This validation data is held aside from other training data to emulate a situation where it has never been exposed to the model [61]. The validation data is usually a randomly selected subset of the whole training population (e.g., 10% of the whole training population, selected randomly). In extreme cases where a dataset is very small it can even be only one instance (i.e., Leave One Out validation [62]).

Model training is highly automated with the exception of hyper-parameter tuning, which is the process for finding the best hyper-parameter combinations. Hyper-parameters are parameters used in ML models that needs to be set before the training process in order to modify model behaviours. Finding the best hyper-parameter combination for a problem can be complicated as there are infinitely many combinations. Historically, ML experts would manually create several combinations for the model to train with and identify the best combination among them. This method, however, is inefficient and vulnerable to human bias. To avoid this, ML experts now use grid-search, random-search and Bayesian optimisation [63–65] for this purpose. It is essential that this process is only performed on the training set. Changes in hyper-parameters can affect results significantly as we can see in the comparison of Figures Fig. 2(c) and Fig. 2(d) where the first is tuned and the second has all but one hyper-parameter manually assigned.

When evaluating the final model, a test score is calculated on the test set (see Fig. 1). Test scores are generally lower than validation scores because they are completely hidden from the model optimisation process and subject to randomness. The test score is a representation on how a model performs when encountered with random unseen data.

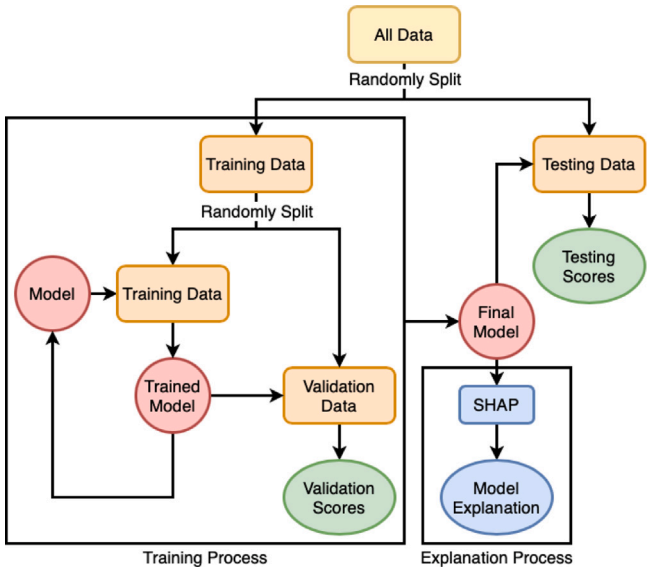


Fig. 1. Data flow of a training, validating, and testing process. SHAP explanation process is not part of the model optimisation and acts as an external component tool specifically for model explanation. It is also illustrated to share its position in the pipeline.

The combination of the test set and validation set can indicate the presence of any model overfitting [66], which occurs when the model fits to the training set so well that it loses its generality on the whole population. This occurs when the model is too sophisticated and starts to fit to the noise. By reporting the validation score after every optimisation step, overfitting can be identified when the validation loss and testing loss start to diverge. In biomarker discovery, classification ability is usually measured by AUC-ROC scores because of its semantics.

Being human-centred and highly case-dependent, explainability is hard to capture by mathematical formulae. It is highly subjective and the utility depends on the application [67]. The most active area of XAI research focuses on post-hoc explanations [68,69] that provide partial quantification on this requirement. A variety of visualisation methods have also been developed to assist users [70].

Shapley values, mentioned above, can be used to determine the relative marginal contribution of a variable towards an individual decision by a ML model (see Appendix A). One of the advantages of Shapley analysis is the level of specificity it affords. SHAP assigns an importance score to individual variables, each belonging to a specific biomarker for each and every patients. In addition to explaining the ML model and why it performs in a certain way, this introduces the possibility of explaining individual measurements and potentially informing treatment plans. These explanations do not, however, relate to validity of the model or have semantics themselves. They are only meaningful when joined with medical knowledge, where invalid contributions could also be observed by clinicians. This is also particularly relevant to personalised medicine [71].

While there is a broad range of literature and techniques for explaining the results or outputs of models including LIME [72], Shapley values and SHAP [38], counterfactual explanations [73] and many more [74–76], Shapley values have a strong theoretical foundation, are model-agnostic, and satisfy key properties of human intuition and reasoning. Shapley based approaches are becoming one of the most widely used explanation techniques and have seen application in many domains [77,78] including healthcare [79].

In our case study, the dataset was split into a training and test set with 80% and 20% of the total data population respectively. All data

Table 1
p-value and effect size (Cohen's *d*) for variables (see Table B.4) ranked top 5 by SHAP.

Ranking	Variable Name	*p*-value	Cohen's *d*
1	panelb_ca125	6.3053×10^{-10}	1.0136
2	panelb_he4	3.7194×10^{-15}	1.3069
3	panelc_ca125_log	1.1866×10^{-11}	1.3287
4	panelc_ca125	1.1866×10^{-11}	0.8452
5	panelc_he4_log	4.7180×10^{-9}	1.0269

Table 2
p-value and effect size for the least important five variables (see Table B.4) as ranked by SHAP.

Ranking	Variable Name	*p*-value	Cohen's *d*
16	panelb_ca15_3	0.0001	0.6477
17	paneld_hepc	0.4455	0.0397
18	panele_mmp_3	0.1562	−0.0285
19	paneld_apo	0.2936	−0.0121
20	paneld_tt	0.4154	0.0018

Table 3
Comparison of rankings generated by Shapley values and *p*-values. See Table B.4 for variable names.

Ranking by Shapley values	1	2	3	4	5	6	7	8
Ranking by *p*-values	5	1	3	4	6	2	7	8

pre-processing and training was done purely on training set, with only the final testing carried out on test set. A random search optimisation technique and 10-fold cross-validation was used in both decision tree and random forest training to ensure stability and optimal performance. Classification ability was measured using AUC-ROC score and model explanation was generated using SHAP with both local and global fidelity.

2.4. Multi-label alternative

In this case study, only binary labels were considered. However, if multiple labels exist (e.g., considering different stages or different subtypes), minor adjustments needs to be made. This includes, but are not limited to, collecting more data, using multi-class classifiers, and using adjusting evaluation metrics as a multi-class AUC-ROC score requires binarizing the output. SHAP could be used in a multi-class scenario but interpreting it becomes more computationally challenging.

3. Result

Code used for generating these results can be found at: https://github.com/jacobvons/PLCO-paper-data-analysis.

3.1. Statistical analysis

In this study, ML provided an extremely effective tool to classify ovarian cancer with statistically significant gains against its algorithmic benchmarks. We used the results from conventional statistical tests as a baseline to highlight the efficacy of our ML methods.

We set the null hypothesis (H_0) as there being no significant difference in the distribution of the variable values as observed on the "healthy" and "case" cohort in the dataset. Type I error rate (α) of the test was set to be 5% following statistical conventions, and reported as-is with no adjustment to α for multiple comparisons. We used Mann–Whitney U test [80] to verify the results generated by SHAP. By calculating the *p*-value and effect sizes as Cohen's *d* [81] for the top-5 ranked variables as shown in Table 1; a conventional rule is to consider a Cohen's *d* of 0.2 as small, 0.5 as medium, and 0.8 as large [82]. The 4th ranked variable had a large effect size while the other top-5 ranked variables had extremely large effect size. This means that the effects of these variables on distinguishing "healthy" and "case" instances in the dataset were extremely strong and significant.

To further validate the ranking of the variables, we compared *p*-values and effect sizes of the top-ranked and the bottom-ranked variables (see Table 2). To conclude, the least important variables had significantly larger *p*-values and much smaller effect sizes than those of the most important variables. This means that the ranking was meaningful and consistent from a statistical perspective.

A comparison of FI ranking generated by the SHAP values and *p*-values was measured using the Wilcoxon Signed Rank test. There was no statistically significant difference between the two rankings, with a *p*-value of 0.97, meaning SHAP values generated FI profile was valid when compared with previous methods. Clear similarity in the ranking was observed especially in the top-eight ranked variables (see Table 3). Differences in the ranking are due to the difference between the two methods, SHAP values take intra-variable interactions into consideration when calculating contribution while *p*-values only measure the impact of single variables. This also means that statistically insignificant variables can be considered important when measured using the SHAP values.

A statistical advisor from the Australian National University Statistical Support Network (SSN) was involved in both the analysis design and outcome reporting.

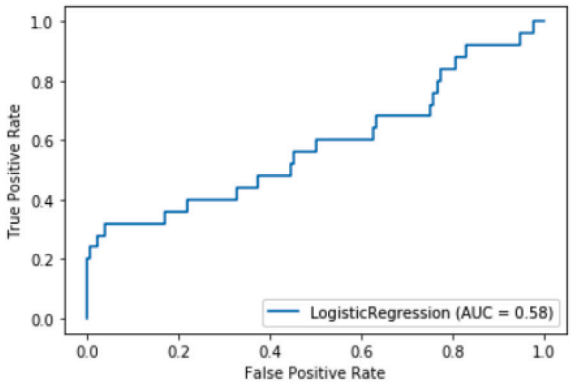
3.2. Result evaluation

Evaluating the result is essential for obtaining a holistic view of the model performance. Results from the three models (logistic regression, decision tree, and random forest) were evaluated from classification ability and explainability perspectives to mimic a real application scenario. Testing results of the three models are shown by the ROC in Figures Fig. 2(a), Fig. 2(b), and Fig. 2(c). It is straightforward to see from the AUC that the random forest exhibits the best predictive behaviour. In a real world application, if a researcher finds this accuracy acceptable, the next step is to understand the predicting process and make decisions based on domain knowledge.

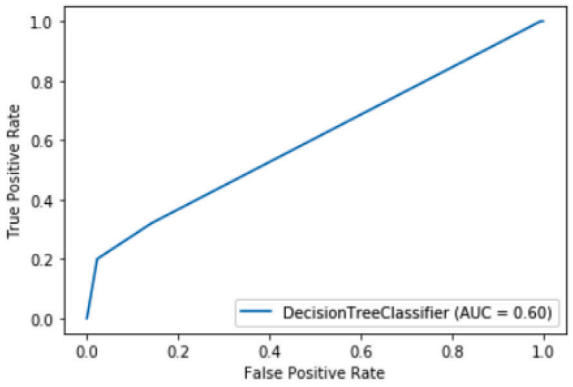
Both logistic regression and decision trees are interpretable. Weights in logistic regression represent the contribution of specific variables to the result. Relative FI of variables can be obtained by analysing the weights, where greater absolute value means more contribution. Nodes in decision trees represent its splitting conditions. The earlier a variable appears in its tree structure, the more important it is in splitting the classes.

To explain the random forest, we used SHAP to calculate variable attributions with both local and global fidelity. Fig. C.5 provides a global view of the random forest in this case study. Variables such as CA-125, HE4 and their statistical variants are ranked high in Fig. C.5 (see Table B.4 for variables), meaning these variables are most important in distinguishing healthy participants from patients. The ROMA algorithm using these two biomarkers is the state-of-the-art algorithm for assessing the likelihood of malignancy and need for surgery of an adnexal mass, specifically from an ovarian cancer perspective [18,20,21]. This further validates the effectiveness of the approach using SHAP.

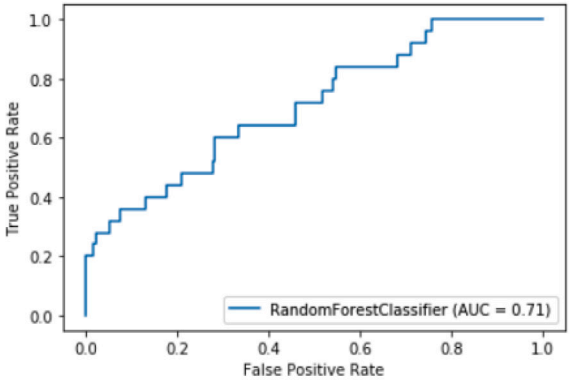
Moreover, relationships between the variables and the SHAP values were obtained by analysing the result from Fig. C.5 where dots are data instances. For example, panelb_ca125 had many SHAP values well below model average, meaning those instances contribute negatively to the predicted value. The same instances also posed high variable values. This means a high panelb_ca125 value contributes towards a negative classification result. For the same variable, we could also see high values contributing positively to the result, which seems contradictory. These controversial instances and their implications to the classification could be further analysed with local fidelity on an instance level. There could be something special about these particular participants.



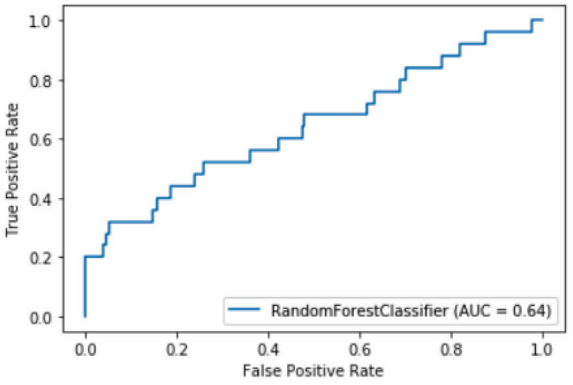
(a) Classification result of the logistic regression. AUC-ROC=0.58



(b) Classification result of the decision tree. AUC-ROC=0.6



(c) Classification result of the random forest. AUC-ROC=0.71



(d) Classification result of the random forest with max depth parameter not optimised. AUC-ROC=0.64

Fig. 2. Comparison between different AUC-ROC.

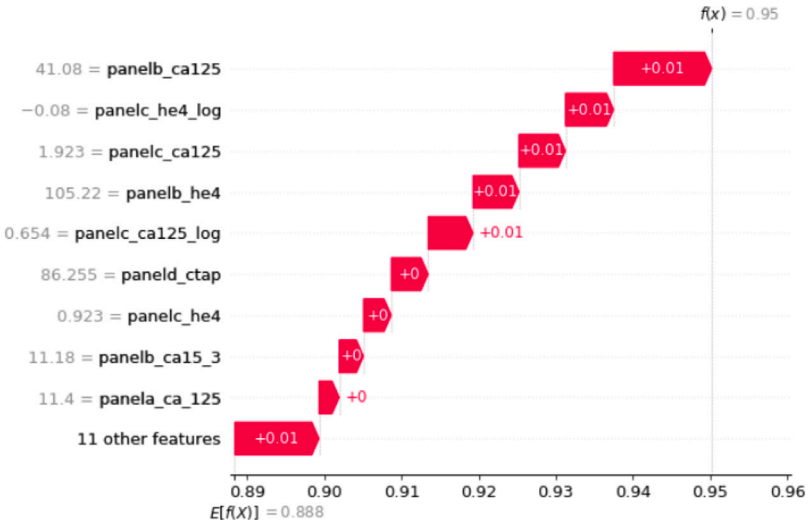


Fig. 3. Training instance one. Probability of being class non-case is 95%, ground truth of the instance is "non-case".

With a local fidelity, SHAP values identify contributions to classifications made on specific instances and variables to highlight uniqueness. This potentially explained why specific variables are different. For example, the variable contributions to the model classification of the first and second instance in the training set were compared in Figs. 3 and 4.

In the case of the first training instance, the top nine variables were all positive but provided only small contribution to the classification. With the variable values shown on the left, we conclude that during the training phase these variables contributed towards a correct classification. Relying simply on this one instance might not provide a holistic view. In Fig. 4, an elevated value of CA-125, as

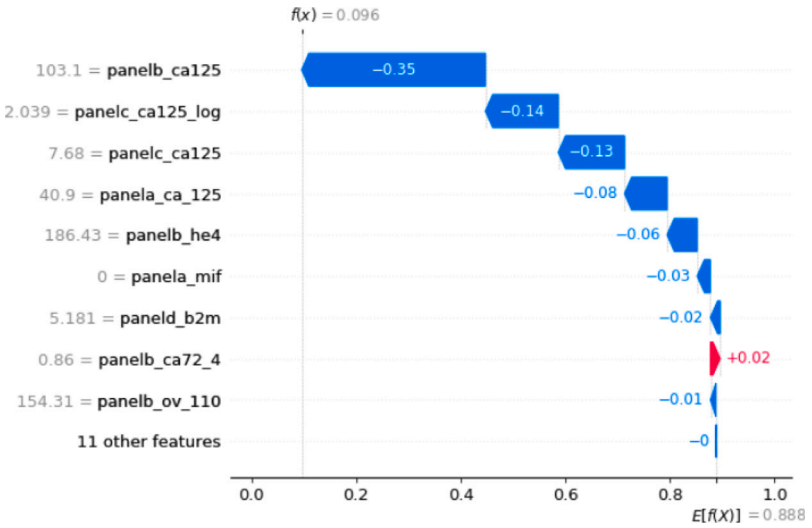


Fig. 4. Training instance two. Probability of being class non-case is 9.6%, ground truth of the instance is "case".

shown in the top two rows, had a significant contribution towards the classification of and instance being a positive case, which was correct. CA-125 contributed most towards the correct classification in the second instance as well. Different values of CA-125 changed the contribution significantly, for example, an elevated value consistently pushed the classification towards being a "case". Based on this, a confident conclusion is that CA-125 is an important biomarker in the classification of ovarian cancer. Similarly, when combined with HE4, the class was often closer to the ground truth, meaning it was preferable to use these two variables together when making a classification, which aligned with domain knowledge of ROMA having better behaviour than CA-125 alone [18,20,21]. Such alignment of results provided convincing evidence that the underlying ML method can be trusted. It could also inform a personalised treatment plan, where the influence of certain biomarkers in certain patients is stronger or weaker.

In summary, the case study demonstrated a workflow from data collection to result evaluation through using classification and XAI. During the process, we discussed, explained, and justified key decisions in the pipeline that help make the result more explainable. To evaluate the results, we gave examples of some of the ways SHAP results can be interpreted and utilised in combination with domain knowledge. This pipeline was effective, general, and applicable to any similar tabular set of biomarkers.

4. Conclusion

In this paper, we outlined a general ML pipeline combining conventional machine learning with Shapley analysis as an explanation tool for the biomarker discovery tasks in the medical domain. With a case study using the PLCO Ovarian Biomarkers dataset, we demonstrated effectiveness of the pipeline as well as its consistency with conventional statistical analysis result and agreement with established domain knowledge.

While many of the issues with explainable techniques [68] are also relevant to Shapley values, such as the interpretability gap and the locality problem, they are much less relevant in the context of this study. We advocate for the feature explanations to be presented at the end of the ML pipeline to be used in conjunction with all other relevant aspects of the given study. This information, just like any other domain information, needs to be carefully inspected and challenged [83]. While there are issues associated with Shapley values, there is significant research progress regarding the improvement of existing techniques [84,85].

One limitation of using Shapley analysis is the higher computational cost, especially on big cohorts with a large number of variables. XAI will take longer than conventional statistical tests, but the time cost can be mitigated by using dimensionality reduction methods as well as special Shapley value approximation algorithms (e.g., SHAP).

A major advantage of Shapley analysis compared to conventional statistical tests is the enhanced explainability of the results. Closely related to accountability issues in medical science, it provides an extra layer of trust to domain experts. When combined with medical data, Shapley values enhance the utility of ML methods for hypothesis generation in addition to hypothesis testing [6]. There is also some evidence to suggest that such explanations inspire some degree of understanding, awareness, and trust, particularly for those with domain knowledge in the given task [86,87].

Challenges also arise from several aspects of XAI in medical applications. First, Shapley values provide feature importance but not knowledge. Proper domain knowledge needs to be associated with the Shapley values to gain insights. Second, there is currently no universal definition of explainability across or within domains. Evaluating the outcome is still highly human-centric, which brings risks and instability into the system. New metrics are required to definitively verify "true intentions" of ML models with explanation provided by XAI. Further, Shapley analysis infers correlation but not causal relationships between variables and labels, which makes the "true intention" analysis more important. Finally, it is also worth noting that Shapley analysis is a post-hoc analysis tool, meaning it would not improve the model classification ability and should only be used to explain a diagnosis after it is made.

In this study, we developed a simplified ovarian cancer diagnosis model using binary labels. In reality, there are multiple stages and types of the disease. Currently, early detection, focusing on stage I and II, is the most immediate concern in clinical practices. Also, although epithelial ovarian cancer is the most common sub-type, other sub-types including germ cell tumours and sex cord stromal tumours are as fatal. It is essential even at the initial stage of data collection that researchers are aware of these subtypes.

It is worth noting that the results from this study have not been compared with those from existing literature generated from methods such as ROMA. Some meta-analysis has shown that ROMA has specificity and sensitivity as high as more than 90% [88–90] while in this study, AUC-ROC score for an optimised random forest classifier is only 71%. However, no ROMA results have been reported for the PLCO Ovarian dataset so a direct comparison is not possible. Since ROMA is a deterministic method while a ML approach depends on randomised and repeated experiments, dataset quality, including (but not limited to) sample size and sample generality provide added explanations unavailable to ROMA.

Table B.4
Top 20 most important variables as selected by the random forest classifier.

Ranking	Variable Name	FI	Description
1	panelb_ca125	0.09998	CA-125 (U/mL)
2	panelc_ca125	0.05858	CA-125 (Fold Change)
3	panelb_he4	0.05183	HE4 (nM)
4	panelc_ca125_log	0.05000	CA125 (Log Transformed Fold Change)
5	panelc_he4	0.04486	HE4 (Fold Change)
6	panelc_he4_log	0.03934	HE4 (Log Transformed Fold Change)
7	panela_ca_125	0.03618	CA-125 (IU/ml)
8	paneld_hepc	0.03132	Hepcidin-25 (TIC)
9	panela_mif	0.02848	MIF (pg/ml)
10	panelb_ov_110	0.02671	OV-110 (pg/mL)
11	panelb_klk6	0.02560	KLK6 (ug/L)
12	panelc_spondin2_log	0.02547	Spondin2 (Log Transformed Fold Change)
13	panelb_ca72_4	0.02478	CA72.4 (U/mL)
14	panelb_ca15_3	0.02283	CA15.3 (U/mL)
15	panelc_slpi	0.02215	SLPI (Fold Change)
16	paneld_tt	0.02102	Transthyretin (TIC)
17	paneld_ctap	0.02075	CTAPIII (TIC)
18	panele_mmp_3	0.02039	MMP-3 (Scaled Fluorescent Intensity)
19	paneld_b2 m	0.01998	Beta-2-microglubulin (TIC)
20	paneld_apo	0.01960	Apolipoprotein A-I (TIC)

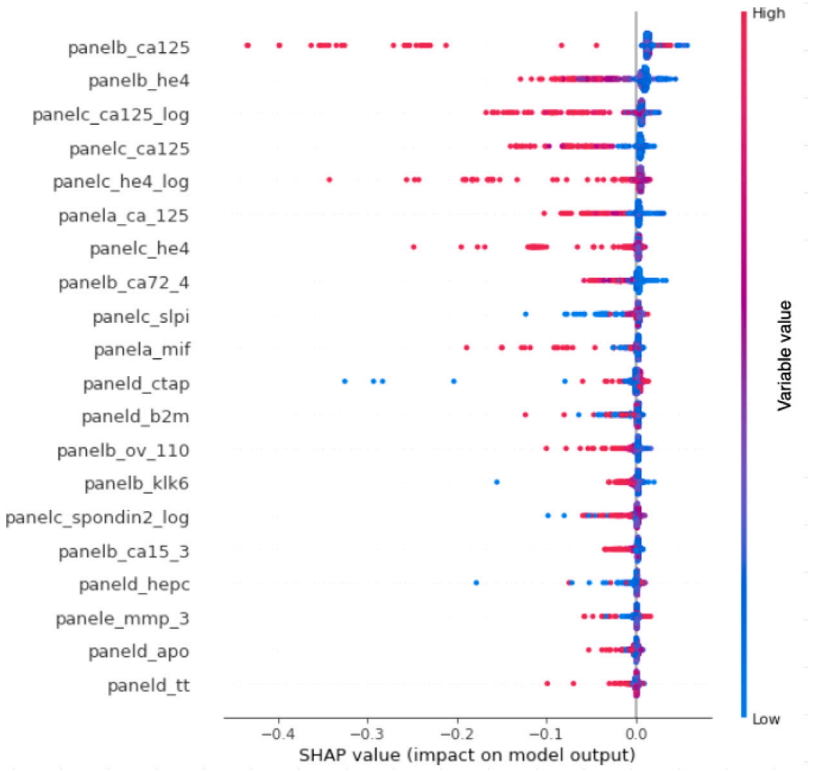


Fig. C.5. Global variable attribution and FI ordering using SHAP. The difference of ranking compared with Table B.4 is caused by different measurement, where Table B.4 relies on inherent training mechanism (e.g., gini-index or impurity reduction) and this plot uses Shapley values. Similarity is related to the contribution towards a good separation.

In general, this approach has broader implications for both computer and health sciences. For the medical community, this work provides a clear guideline for adopting ML methods in research studies, especially when accountability issues are an imperative. For the computing field, this work highlights areas that would benefit from improvements, such as reducing resource requirements to make SHAP faster and more efficient. Working together with best practices, medical domain experts and computer scientists may be able to advance the discovery of novel disease biomarkers through the use of explainable models.

CRediT authorship contribution statement

Weitong Huang: Conceived and designed the experiments and analysis, Collected the data, Performed the model building and data analysis, Took the lead in writing the manuscript, Provided critical feedback and helped shape the research, analysis and manuscript. **Hanna Suominen:** Conceived and designed the experiments and analysis, Provided critical feedback and helped shape the research, analysis and manuscript. **Tommy Liu:** Conceived and designed the experiments and analysis, Provided support in interpreting the results, Provided critical feedback and helped shape the research, analysis and

manuscript. **Gregory Rice:** Provided medical domain expertise in background and analysis, Provided critical feedback and helped shape the research, analysis and manuscript. **Carlos Salomon:** Provided medical domain expertise in background and analysis, Provided critical feedback and helped shape the research, analysis and manuscript. **Amanda S. Barnard:** Conceived and designed the experiments and analysis, Provided critical feedback and helped shape the research, analysis and manuscript.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competng interests: Carlos Salomon Gallo reports financial support was provided by National Health and Medical Research Council. Carlos Salomon Gallo reports financial support was provided by Australian Government Department of Health and Aged Care.

Acknowledgements

This study is supported by the Medical Research Future Fund, Australia under award numbers MRF1199984 (CS) and GA187319 (CS, GER, HS and AB), the National Health and Medical Research Council, Australia under award number NHMRC 1195451 (CS), The Lion Medical Research Foundation 2015001964 (CS), The Donald & Joan Wilson Foundation Ltd 2020000323 (CS), and Ovarian Cancer Research Foundation 2018001167 (CS).

Appendix A. SHAP

For a particular variable i on a particular model f, its Shapley value, ϕ_i, can be computed as variable attributions for the weighted average for all possible variable subsets S:

$$\phi_i = \sum_{S \subset F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S\left(x_S\right) \right] \tag{A.1}$$

where f_S is a model trained without the particular feature i, $f_{S\cup\{i\}}$ is a model trained with feature i, and x_S, $x_{S\cup\{i\}}$ are the values of the feature inputs. Compared to calculation using definition of Shapley values directly, a less computationally intense and more widely used framework is SHAP, described in [38].

Appendix B. Variable importance profile

See Table B.4.

Appendix C. Global SHAP

See Fig. C.5.

References

[1] I.J. Jacobs, H. Rivera, D.H. Oram, R.C. Bast Jr., Differential diagnosis of ovarian cancer with tumour markers CA 125, CA 15-3 and TAG 72.3, BJOG: Int. J. Obstetr. Gynaecol. 100 (12) (1993) 1120–1124.
[2] I.J. Jacobs, S. Skates, A.P. Davies, R.P. Woolas, A. Jeyerajah, P. Weidemann, K. Sibley, D.H. Oram, Risk of diagnosis of ovarian cancer after raised serum CA 125 concentration: A prospective cohort study, Bmj 313 (7069) (1996) 1355–1358.
[3] I.J. Jacobs, S.J. Skates, N. MacDonald, U. Menon, A.N. Rosenthal, A.P. Davies, R. Woolas, A.R. Jeyarajah, K. Sibley, D.G. Lowe, et al., Screening for ovarian cancer: A pilot randomised controlled trial, Lancet 353 (9160) (1999) 1207–1210.
[4] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260.
[5] M. Chen, S. Mao, Y. Liu, Big data: A survey, Mob. Netw. Appl. 19 (2) (2014) 171–209.
[6] C.H. Lee, H.-J. Yoon, Medical big data: Promise and challenges, Kidney Res. Clin. Practice 36 (1) (2017) 3.
[7] D.V. Dimitrov, Medical internet of things and big data in healthcare, Healthcare Inform. Res. 22 (3) (2016) 156–163.
[8] L. Zhang, H. Wang, Q. Li, M.-H. Zhao, Q.-M. Zhan, Big data and medical research in China, Bmj 360 (2018).
[9] A. Garg, V. Mago, Role of machine learning in medical research: A survey, Comp. Sci. Rev. 40 (2021) 100370.
[10] X. Ni, W. Ouyang, H. Jeong, J.-T. Kim, A. Tzaveils, A. Mirzazadeh, C. Wu, J.Y. Lee, M. Keller, C.K. Mummidisetty, et al., Automated, multiparametric monitoring of respiratory biomarkers and vital signs in clinical and home settings for COVID-19 patients, Proc. Natl. Acad. Sci. 118 (19) (2021).
[11] T.H. Rim, G. Lee, Y. Kim, Y.-C. Tham, C.J. Lee, S.J. Baik, Y.A. Kim, M. Yu, M. Deshmukh, B.K. Lee, et al., Prediction of systemic biomarkers from retinal photographs: Development and validation of deep-learning algorithms, Lancet Digit. Health 2 (10) (2020) e526–e536.
[12] O. Jones, R. Matin, M. van der Schaar, K.P. Bhayankaram, C. Ranmuthu, M. Islam, D. Behiyat, R. Boscott, N. Calanzani, J. Emery, et al., Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: A systematic review, Lancet Digit. Health 4 (6) (2022) e466–e476.
[13] M. Terashima, T. Irino, Predicting peritoneal recurrence by artificial intelligence, Lancet Digit. Health 4 (5) (2022) e293–e294.
[14] M.A. Robb, P.M. McInnes, R.M. Califf, Biomarkers and surrogate endpoints: developing common terminology and definitions, JAMA 315 (11) (2016) 1107–1108.
[15] R.M. Califf, Biomarker definitions and their applications, Exp. Biol. Med. 243 (3) (2018) 213–221.
[16] M.S. Pepe, R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist, M. Winget, Y. Yasui, Phases of biomarker development for early detection of cancer, J. Natl. Cancer Inst. 93 (14) (2001) 1054–1061.
[17] E. Moss, J. Hollingworth, T. Reynolds, The role of CA125 in clinical practice, J. Clin. Pathol. 58 (3) (2005) 308–312.
[18] V. Dochez, H. Caillon, E. Vaucel, J. Dimet, N. Winer, G. Ducarme, Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review, J. Ovarian Res. 12 (1) (2019) 1–9.
[19] S. Ferraro, F. Braga, M. Lanzoni, P. Boracchi, E.M. Biganzoli, M. Panteghini, Serum human epididymis protein 4 vs carbohydrate antigen 125 for ovarian cancer diagnosis: A systematic review, J. Clin. Pathol. 66 (4) (2013) 273–281.
[20] C. Romagnolo, A.E. Leon, A.S. Fabricio, M. Taborelli, J. Polesel, L. Del Pup, A. Steffan, S. Cervo, A. Ravaggi, L. Zanotti, et al., HE4, CA125 and risk of ovarian malignancy algorithm (ROMA) as diagnostic tools for ovarian cancer in patients with a pelvic mass: An Italian multicenter study, Gynecol. Oncol. 141 (2) (2016) 303–311.
[21] M. Montagnana, E. Danese, O. Ruzzenente, V. Bresciani, T. Nuzzo, M. Gelati, G.L. Salvagno, M. Franchi, G. Lippi, G.C. Guidi, The ROMA (risk of ovarian malignancy algorithm) for estimating the risk of epithelial ovarian cancer in women presenting with pelvic mass: Is it really useful? Clin. Chem. Lab. Med. 49 (3) (2011) 521–525.
[22] L. Nguyen, S.J. Cardenas-Goicoechea, P. Gordon, C. Curtin, M. Momeni, L. Chuang, D. Fishman, Biomarkers for early detection of ovarian cancer, Women's Health 9 (2) (2013) 171–187.
[23] C. Stewart, C. Ralyea, S. Lockwood, Ovarian cancer: An integrated review, in: Seminars in Oncology Nursing, Vol. 35, (2) Elsevier, 2019, pp. 151–156.
[24] M. Sahakyan, Z. Aung, T. Rahwan, Explainable artificial intelligence for tabular data: A survey, IEEE Access 9 (2021) 135392–135422.
[25] A. Shaban-Nejad, M. Michalowski, J.S. Brownstein, D.L. Buckeridge, Guest editorial explainable AI: Towards fairness, accountability, transparency and trust in healthcare, IEEE J. Biomed. Health Inf. 25 (7) (2021) 2374–2375.
[26] A.J. London, Artificial intelligence and black-box medical decisions: Accuracy versus explainability, Hastings Center Report 49 (1) (2019) 15–21.
[27] E. Vayena, A. Blasimme, I.G. Cohen, Machine learning in medicine: addressing ethical challenges, PLoS Med. 15 (11) (2018) e1002689.
[28] E. Halilaj, A. Rajagopal, M. Fiterau, J.L. Hicks, T.J. Hastie, S.L. Delp, Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities, J. Biomech. 81 (2018) 1–11.
[29] S. Reddy, Explainability and artificial intelligence in medicine, Lancet Digit. Health 4 (4) (2022) e214–e215.
[30] M. Ghassemi, T. Naumann, P. Schulam, A.L. Beam, I.Y. Chen, R. Ranganath, A review of challenges and opportunities in machine learning for health, AMIA Summits Transl. Sci. Proc. 2020 (2020) 191.
[31] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2020) 4793–4813.
[32] L.L. Vercio, K. Amador, J.J. Bannister, S. Crites, A. Gutierrez, M.E. MacDonald, J. Moore, P. Mouches, D. Rajashekar, S. Schimert, et al., Supervised machine learning tools: A tutorial for clinicians, J. Neural Eng. 17 (6) (2020) 062001.
[33] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, Vol. 398, John Wiley & Sons, 2013.
[34] Y.-Y. Song, L. Ying, Decision tree methods: Applications for classification and prediction, Shanghai Arch. Psychiatry 27 (2) (2015) 130.
[35] S.-C. Wang, Artificial neural network, in: Interdisciplinary Computing in Java Programming, Springer, 2003, pp. 81–100.
[36] Y. Qi, Random forest for bioinformatics, in: Ensemble Machine Learning, Springer, 2012, pp. 307–323.
[37] T. Dietterich, Overfitting and undercomputing in machine learning, ACM Comput. Surv. 27 (3) (1995) 326–327.

[38] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[39] D. Goldmann, System failure versus personal accountability—the case for clean hands, N. Engl. J. Med. 355 (2) (2006) 121–123.
[40] P. Cheung, Public Trust in Medical Research?: Ethics, Law and Accountability, CRC Press, 2018.
[41] A. Kiseleva, AI as a medical device: Is it enough to ensure performance transparency and accountability in healthcare? Eur. Pharm. Law Rev. (1) (2020).
[42] C.-W. Park, S.W. Seo, N. Kang, B. Ko, B.W. Choi, C.M. Park, D.K. Chang, H. Kim, H. Kim, H. Lee, et al., Artificial intelligence in health care: Current applications and issues, J. Korean Med. Sci. 35 (42) (2020).
[43] L. Felländer-Tsai, AI ethics, accountability, and sustainability: Revisiting the hippocratic oath, Acta Orthopaedica 91 (1) (2020) 1–2.
[44] I. de Miguel, B. Sanz, G. Lazcoz, Machine learning in the EU health care context: exploring the ethical, legal and social issues, Inf., Commun. Soc. 23 (8) (2020) 1139–1153.
[45] J.K. Gohagan, P.C. Prorok, R.B. Hayes, B.S. Kramer, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Project Team, The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the national cancer institute: history, organization, and status, Control Clin. Trials 21 (6) (2000) 251S–272S.
[46] H. Wang, C. Ma, L. Zhou, A brief review of machine learning and its application, in: 2009 International Conference on Information Engineering and Computer Science, IEEE, 2009, pp. 1–4.
[47] K.R. Foster, R. Koprowski, J.D. Skufca, Machine learning, medical diagnosis, and biomedical engineering research-commentary, Biomed. Eng. Online 13 (1) (2014) 1–9.
[48] Y. Roh, G. Heo, S.E. Whang, A survey on data collection for machine learning: A big data - AI integration perspective, IEEE Trans. Knowl. Data Eng. 33 (4) (2021) 1328–1347.
[49] C.M. Coticchia, J. Yang, M.A. Moses, Ovarian cancer biomarkers: Current options and future promise, J. Natl. Compr. Cancer Netw. 6 (8) (2008) 795–802.
[50] S. Kato, L. Abarzua-Catalan, C. Trigo, A. Delpiano, C. Sanhueza, K. García, C. Ibañez, K. Hormazábal, D. Diaz, J. Brañes, et al., Leptin stimulates migration and invasion and maintains cancer stem-like properties in ovarian cancer cells: An explanation for poor outcomes in obese women, Oncotarget 6 (25) (2015) 21100.
[51] M. Schmitt, V. Magdolen, F. Yang, M. Kiechle, J. Bayani, G.M. Yousef, A. Scorilas, E.P. Diamandis, J. Dorn, Emerging clinical importance of the cancer biomarkers kallikrein-related peptidases (KLK) in female and male reproductive organ malignancies, Radiol. Oncol. 47 (4) (2013) 319.
[52] I. Hellström, J. Raycraft, M. Hayden-Ledbetter, J.A. Ledbetter, M. Schummer, M. McIntosh, C. Drescher, N. Urban, K.E. Hellström, The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma, Cancer Res. 63 (13) (2003) 3695–3700.
[53] A. Tamir, U. Jag, S. Sarojini, C. Schindewolf, T. Tanaka, R. Gharbaran, H. Patel, A. Sood, W. Hu, R. Patwa, et al., Kallikrein family proteases KLK6 and KLK7 are potential early detection and diagnostic biomarkers for serous and papillary serous ovarian cancer subtypes, J. Ovarian Res. 7 (1) (2014) 1–15.
[54] F. Su, K.R. Kozak, S. Imaizumi, F. Gao, M.W. Amneus, V. Grijalva, C. Ng, A. Wagner, G. Hough, G. Farias-Eisner, et al., Apolipoprotein AI (apoa-i) and apoa-i mimetic peptides inhibit tumor development in a mouse model of ovarian cancer, Proc. Natl. Acad. Sci. 107 (46) (2010) 19997–20002.
[55] B. Gericke, J. Raila, J. Sehouli, S. Haebel, D. Könsgen, A. Mustea, F.J. Schweigert, Microheterogeneity of transthyretin in serum and ascitic fluid of ovarian cancer patients, BMC Cancer 5 (1) (2005) 1–9.
[56] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[57] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, Ieee, 2015, pp. 1200–1205.
[58] P. Smialowski, D. Frishman, S. Kramer, Pitfalls of supervised feature selection, Bioinformatics 26 (3) (2010) 440–443.
[59] U. Pawar, D. O'Shea, S. Rea, R. O'Reilly, Incorporating explainable artificial intelligence (XAI) to aid the understanding of machine learning in the healthcare domain, in: AICS, 2020, pp. 169–180.
[60] G. Aihemaiti, X. Zhang, T. Zhang, K. Hu, J. Xu, Z. Zhang, Q. Zhao, N. Song, F. Liu, Y. Yang, et al., Development and validation of a nomogram to predict the risk of major adverse cardiovascular events in patients with non-ST-segment elevation myocardial infarction, 2022.
[61] J.M. Twomey, A.E. Smith, Validation and verification, in: Artificial Neural Networks for Civil Engineers: Fundamentals and Applications, ASCE Press, New York, 1997, pp. 44–64.
[62] T.-T. Wong, Performance evaluation of classification algorithms by K-fold and leave-one-out cross validation, Pattern Recognit. 48 (9) (2015) 2839–2846.
[63] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2) (2012).
[64] P. Liashchynskyi, P. Liashchynskyi, Grid search, random search, genetic algorithm: A big comparison for NAS, 2019, arXiv preprint arXiv:1912.06059.
[65] T.T. Joy, S. Rana, S. Gupta, S. Venkatesh, Hyperparameter tuning for big data using Bayesian optimisation, in: 2016 23rd International Conference on Pattern Recognition, ICPR, IEEE, 2016, pp. 2574–2579.
[66] D. Berrar, Cross-validation, 2019.
[67] M. Ribera, A. Lapedriza, Can we do better explanations? A proposal of user-centered explainable AI, in: IUI Workshops, Vol. 2327, 2019, p. 38.
[68] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, Lancet Digit. Health 3 (11) (2021) e745–e750.
[69] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Inf. Fusion 77 (2022) 29–52.
[70] M. Pocevičiūtė, G. Eilertsen, C. Lundström, Survey of XAI in digital pathology, in: Artificial Intelligence and Machine Learning for Digital Pathology, Springer, 2020, pp. 56–88.
[71] G.D. Kitsios, D.M. Kent, Personalised medicine: Not just in our genes, BMJ 344 (2012).
[72] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[73] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv. JL & Tech. 31 (2017) 841.
[74] M.T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, (1) 2018.
[75] B. Kim, R. Khanna, O.O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, Adv. Neural Inf. Process. Syst. 29 (2016).
[76] C. Molnar, Interpretable Machine Learning, Lulu. com, 2020.
[77] Y. Nohara, K. Matsumoto, H. Soejima, N. Nakashima, Explanation of machine learning models using shapley additive explanation and application for real data in hospital, Comput. Methods Programs Biomed. 214 (2022) 106584.
[78] R. Rodríguez-Pérez, J. Bajorath, Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions, J. Comput. Aided Mol. Des. 34 (2020) 1013–1026.
[79] L. Ibrahim, M. Mesinovic, K.-W. Yang, M.A. Eid, Explainable prediction of acute myocardial infarction using machine learning and shapley values, IEEE Access 8 (2020) 210410–210417.
[80] P.E. McKnight, J. Najab, Mann-whitney u test, in: The Corsini Encyclopedia of Psychology, Wiley Online Library, 2010, p. 1.
[81] G.E. Gignac, E.T. Szodorai, Effect size guidelines for individual differences researchers, Personality Individ. Differ. 102 (2016) 74–78.
[82] J.J. McGough, S.V. Faraone, Estimating the size of treatment effects: Moving beyond p values, Psychiatry (Edgmont) 6 (10) (2009) 21.
[83] G. Cinà, T. Röber, R. Goedhart, I. Birbil, Why we do need explainable ai for healthcare, 2022, arXiv preprint arXiv:2206.15363.
[84] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: Quantifying the limits of the Shapley value for explanations, Adv. Neural Inf. Process. Syst. 34 (2021) 26598–26608.
[85] I. Covert, S.-I. Lee, Improving KernelSHAP: Practical Shapley value estimation using linear regression, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3457–3465.
[86] X. Wang, M. Yin, Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 318–328.
[87] Y. Alufaisan, L.R. Marusich, J.Z. Bakdash, Y. Zhou, M. Kantarcioglu, Does explainable artificial intelligence improve human decision-making? in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, (8) 2021, pp. 6618–6626.
[88] R. Cui, Y. Wang, Y. Li, Y. Li, Clinical value of ROMA index in diagnosis of ovarian cancer: meta-analysis, Cancer Manag. Res. 11 (2019) 2545.
[89] J. Wang, J. Gao, H. Yao, Z. Wu, M. Wang, J. Qi, Diagnostic accuracy of serum HE4, CA125 and ROMA in patients with ovarian cancer: a meta-analysis, Tumor Biol. 35 (2014) 6127–6138.
[90] V. Kumar, S. Rajan, S. Gupta, N. Akhtar, S. Sharma, P. Sinha, S. Misra, A. Chaturvedi, Diagnostic value of risk of malignancy algorithm (ROMA) in adnexal masses, J. Obstetr. Gynecol. India 70 (2020) 214–219.