TCR TRANSLATIONAL CANCER RESEARCH
ADVANCES CLINICAL MEDICINE TOWARD THE GOAL OF IMPROVING PATIENTS' QUALITY OF LIFE
JOURNAL CITATION REPORTS® IMPACT FACTOR 1.241 © 2021 Clarivate Analytics

Peer Review File

Article information: https://dx.doi.org/10.21037/tcr-22-1626

Reviewer A

1. The original publication by Li et al did not report on the AUC of the gradient-boosted tree they trained. How can one draw insights from the SHAP plots if the discriminatory ability of the model that SHPA analysis is based on is not measured? Did other studies that used SHAP report on performance metrics of their models?
   - Reply: The publication by Li et al is limited by not reporting model metrics. Although the rationale is not given, presumably it is related to the manuscript primarily aiming to illustrate examples of the utility of the SHAP framework in oncology, specifically for improving interpretability. There is discussion of model validation in the last paragraph of the results and associated appendix. It is also discussed in the limitations section of the paper, where the authors report using Cox regression to validate results. However, indeed, most publications do report some form of performance metrics such as AUC. Nevertheless, it is our opinion that this publication is useful in illustrating more nuanced trends and is worth including.
   - Changes in the text: N/A
2. Can you compare SHAP vs LIME? How many studies have utilized each methodology and what are the advantages of each one.
   - Reply: SHAP is more commonly used in modern papers due to producing more attractive plots. Conversely, LIME is older and more heavily represented in earlier publications. The two frameworks use different algorithms, with nuanced pros and cons that could be the topic of a whole separate paper that is beyond the scope of the present paper, but at their core they have the unified goal of explaining ML models. A simple analogy is LIME is to SHAP as chi-square is to Fisher's exact test. SHAP is likely more precise at the cost of computational inefficiencies that can be prohibitive in certain scenarios. Ultimately which to use will come down to user preference and the exact application. This website provides a decent summary: https://www.dominodatalab.com/blog/shap-lime-python-libraries-part-1-great-explainers-pros-cons
   - Changes in the text: The core goal of both algorithms was added to the third paragraph of the introduction. We have also added Table 1 which summarizes the pros and cons of each algorithm.

Reviewer B

Major comments

1. Through the manuscript, it is emphasized that clinicians have a hard time understanding complicated, elaborate (ML) models. While this is true, it is not only the clinicians, but also the patients (or any other end-user), decision makers, and many other stakeholders that could greatly benefit from a more transparent approach. This should be made clear all through the text
   - Reply: This is an excellent point. Though physicians are likely the primary consumer of most of these models, several other end-users will be exposed.
   - Changes in the text: Throughout the manuscript clinician was updated to end-user and the initial definition of end-user was added to the introduction.
2. L65 – The need for more elaborate complex models needs to be expanded. There are many shortcomings of classical statistical models that need to be, at least, mentioned
   - Reply: Acknowledged
   - Changes in the text: A few examples were added on where ML may be beneficial, including modelling non-linear relationships, interaction effects, or image analysis
3. L67 – The benefits of using more complex models should be explained further, so the reader can better grasp the trade-off between complexity and accuracy
   - Reply: Acknowledged
   - Changes in the text: We have added rationale for using more complex models to the introduction.
4. L81 – Although this isn't a technical paper, the authors should provide at least a short description of how LIME and SHAP work (~1 paragraph for each of them). It would be a pity to use a black box to explain another black box ;) . Moreover, the authors should provide an lengthier explanation on why they decided to focus only on these two approaches, given there are many XAI techniques in the field being used. Of particular interest is the use of SHAP, since it presents several (mathematical) properties that are not present in other methods, something that is definitely worth mentioning.
   - Reply: The query was not limited specifically to SHAP and LIME, though these were emphasized since they are model-agnostic and widely used across ML. However, since SHAP and LIME comprise almost every study, we have refocused the paper to model-agnostic frameworks and removed resources that do not use SHAP or LIME. We do feel full paragraphs explaining SHAP and LIME is beyond the scope of this paper, but do think including a table with key points is a reasonable compromise.

- Changes in the text: We have added a table providing details of SHAP and LIME and removed references that use other XAI frameworks

5. L85 – While this is true, there are a number of studies being published that use XAI in different points of a patient's care path. It is important that the reader identifies that XAI can be applied in pretty much all different steps of it. Actually, I think it would benefit the structure of the manuscript greatly to include a figure of the patient care pathway with the different steps. Then, the authors could follow the same order/structure when providing the different examples
   - Reply: While this is true that XAI can be applied to any point in patient care, it is our opinion that a figure detailing the patient care pathway, which almost all readers will be keenly aware of, would not provide significant value to the understanding of the paper. Our rationale is that the available studies and patient care disciplines do not necessarily have a uniform flow that would lend itself to such a figure. It is our opinion that separating the paper into its respective sections sufficiently addresses this goal.
   - Changes in the text: N/A
6. There are a couple of important works in the field that are missing in the authors' literature review of XAI, such as Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017); Holzinger, Andreas, et al. "What do we need to build explainable AI systems for the medical domain?." arXiv preprint arXiv:1712.09923 (2017).
   - Reply: These papers would not have been included in the literature query since they are not specific to oncology. We did add them to the list of references in the Introduction where XAI is introduced.
   - Changes in the text: These references have been added to the Introduction
7. All examples – The authors should provide a little bit more details on the methods used in the studies, such as what was the goal of the study, what was the prediction target, what models were used and, more importantly, what XAI technique was applied. This can be done in one or two sentences top at the beginning of each paragraph corresponding to a study. Some examples have this, but not all of them. This would allow the reader to better get a grasp of what each study was trying to accomplish.
   - Reply: The details of the studies will be updated with more details
   - Changes in the text: The study paragraphs have been updated.
8. L124 – The work in Jansen, T. et al. Machine learning explainability in breast cancer survival. In Studies in Health Technology and Informatics, vol. 270: Digital Personalized Health and Medicine, 307–311 (2020). is relevant and

should be included as well, since they made a comparison between LIME and SHAP for breast cancer survival.

- Reply: This is a useful study. Thank you.
- Changes in the text: We have added a paragraph on this study.

9. L130 – The fact that XGB doesn't output HRs is not really the reason why XAI is needed. XAI (SHAP in this case) allows generating a mathematically-based reasoning of how a model yielded its output – ranked survival in this case. Moreover, currently there's research being performed on calculating HRs based on SHAP values. See the work Sundrani S, Lu J. Computing the Hazard Ratios Associated With Explanatory Variables Using Machine Learning Models of Survival Data. JCO Clin Cancer Inform. 2021 Mar;5:364–78.
   - Reply: Acknowledged
   - Changes in the text: This statement was clarified. The goal is was suggest that some way of understanding how the output is computed is helpful.
10. L151 – XAI has also been used for studying recurrence
    - Reply: Acknowledged
    - Changes in the text: Survival was changed to oncologic to reflect this
11. L224-230 – It isn't clear how XAI can help reduce overfitting. Moreover, if that's the case, this is something that isn't exclusive to Radiomics
    - Reply: It helps with overfitting by aiding with feature selection and helping remove unnecessary features (see how in the Kha paper AUC improves with fewer features). Of note, out text does not say this is unique to radiomics. This is just an example of where it was used. We also included this as a column in
    - Changes in the text: We have specified that XAI can improve final feature selection. The other text in the paragraph supplements this statement.
12. L255 – The work by Janssen, Femke M., et al. "Using Explainable Machine Learning to Explore the Impact of Synoptic Reporting on Prostate Cancer." Algorithms 15.2 (2022): 49. Is another relevant example of XAI in Pathology, since it could lead to policy change in how pathological reports are filled in.
    - Reply: Thank you. This is a useful study.
    - Changes in the text: We have added a paragraph on this study.
13. L288 – It would be nice to add a few sentences discussing the consequences of over or under treating patients
    - Reply: This topic is not specific to XAI and is well-documented in the oncology literature. We feel further discussion is beyond the scope of this paper.
    - Changes in the text: N/A
14. L365 – There's current work on leveraging XAI for other type of applications, such as computing the HRs using SHAP values (Sundrani S, Lu J. Computing the

Hazard Ratios Associated With Explanatory Variables Using Machine Learning Models of Survival Data. JCO Clin Cancer Inform. 2021 Mar;5:364–78.) or as a feature selection mechanism (Marcílio, Wilson E., and Danilo M. Eler. "From explanations to feature selection: assessing SHAP values as feature selection mechanism." 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2020.). Please make sure that you include these (and maybe more) examples in this section.

- Reply: Given these are not specific to oncology, we feel these are better suited in the future directions
- Changes in the text: These citations have been added to the future directions section

15. L343 – One of the things that I felt was missing from the whole manuscript – particularly in this section – was the use of XAI for analyzing individual patients, like the work by Kobylińska, Katarzyna, et al. "Explainable Machine Learning for Lung Cancer Screening Models." Applied Sciences 12.4 (2022): 1926.
    - Reply: Thank you. This is a valuable addition.
    - Changes in the text: We have added a paragraph on this paper.
16. L365 and L397 – These sections should be rewritten together cohesively into a single one called “DISCUSSION” (incorporating further comments)
    - Reply: Acknowledged.
    - Changes in the text: We have added a discussion
17. L365 – Either here or in L394’s section, it would be very useful for the authors to provide a general overview of the advantages and disadvantages when comparing LIME and SHAP based on their narrative review. I can imagine that when published, many readers would refer to this paper as a starting point for using these tools and this would be valuable information to them.
    - Reply: It is our opinion that this is better suited for the newly included table
    - Changes in the text: We have included the pros and cons of SHAP and LIME in Table 1.
18. L365 – Besides opening the black box, there is active work in using XAI as an auxiliary tool to improve other models, like the work being done by XXX, where the use SHAP to help computing Cox model HRs. This should be at least mentioned in this section as a future direction
    - Reply: Acknowledged
    - Changes in the text: Per response to point 14, this has been added to future directions
19. L370 – I do not agree completely with the statement that there might be a single best method for XAI. I do not think there is a silver bullet and depending on what

research question the model is trying to address, an appropriate XAI needs to be chosen.

- Reply: The claim is not that some XAI is best, merely that some general framework may be unifying for general applications, while other frameworks will still have utility.
- Changes in the text: N/A

20. L380 – The work by Duval, Alexandre. "Explainable artificial intelligence (XAI)." MA4K9 Scholarly Report, Mathematics Institute, The University of Warwick (2019): 1-53. should be discussed here, since they focus on this particular trade off.
    - Reply: This is a good citation for the identified sentence.
    - Changes in the text: We have added this citation
21. L387 – Importantly, when their performance is not that different from more complex (and opaque) models.
    - Reply: Acknowledged
    - Changes in the text: We have added this qualifying statement to this sentence.
22. L391 – Although I agree that XAI is an emerging trend, I think calling it a revolution might be a bit of an overstatement. Overall, the tone of the whole paper should be moderated.
    - Reply: To be clear, revolutions was referring to big data and personalized medicine. However, we can change the word
    - Changes in the text: Revolutions was changed to advances

Minor comments

1. All through the manuscript – et al. should have a dot at the end
    - Reply: Acknowledged
    - Changes in the text: et al has been updated to et al. throughout the manuscript.
2. L51 – Typo
    - Reply: Acknowledged
    - Changes in the text: Presnts was changed to presents
3. Keywords should be enlisted in alphabetical order
    - Reply: Acknowledged
    - Changes in the text: Keywords are now in alphabetical order
4. L62 – Often, the coefficients of models such as linear regression are also used for interpreting a model
    - Reply: Acknowledged
    - Changes in the text: “coefficients, and p-values” was added to the sentence
5. L83 – Medicine à Healthcare
    - Reply: Acknowledged
    - Changes in the text: Medicine was changed to healthcare

6. “Figure” should always be abbreviated in the text (e.g., Fig. 1), unless when at the beginning of a sentence
   - Reply: Acknowledged
   - Changes in the text: “Figure” has been abbreviated throughout the text.
7. L85 – Missing “the”
   - Reply: We could not identify a missing “the”
   - Changes in the text: N/A
8. L85 – Immature à in an early stage
   - Reply: Acknowledged
   - Changes in the text: Immature was changed to in an early stage
9. L90 – The checklist should be cited as a reference, with its corresponding link
   - Reply: This does not appear to be done in other published narrative reviews, as it is not being used as a reference. Therefore, we did not add it as a citation.
   - Changes in the text: N/A
10. Table 1 – Remove vertical lines, they make reading the table harder. Proper column spacing should be enough for clear interpretation. “Search” should not be capitalized.
    - Reply: Table 1 is formatted per the template provided by TCR. Therefore we did not modify it.
    - Changes in the text: N/A
11. L100 (and through the whole of the manuscript) – Change “prognostication” to “prognosis”
    - Reply: Given how prognostication is being used as a verb to mean “producing a prognosis”, “prognostication” is appropriate and was not changed.
    - Changes in the text: N/A
12. L103 – Reference is needed (e.g., SEER dataset)
    - Reply: Acknowledged
    - Changes in the text: NCDB and SEER as examples were added
13. L105 – Remove “in prognosis”
    - Reply: Acknowledged
    - Changes in the text: Prognosis was changed to outcomes. We do feel the phrase is warranted in clarifying what is improved.
14. L110 – reveals --> revealed
    - Reply: Acknowledged
    - Changes in the text: Reveals was changed to revealed
15. L111 – Start new sentence; are --> is
    - Reply: Acknowledged
    - Changes in the text: The sentence has been split into 2 and are has been changed to is.

16. L115 – Additionally, when Gleason score if 8 or higher
    - Reply: Acknowledged
    - Changes in the text: The phrase was changed to "Additionally, when Gleason score is 8 or higher"
17. L117 – differentiate between favorable and unfavorable intermediate risk prostate cancer
    - Reply: Acknowledged
    - Changes in the text: The phrase was changed to "differentiate between favorable and unfavorable intermediate risk prostate cancer"
18. L128 – extreme gradient boosted (XGB) tree algorithm
    - Reply: Acknowledged
    - Changes in the text: The "XGB" was moved
19. L153, L158 – AUC should be reported as a number between 0 and 1, but not as a percentage.
    - Reply: Acknowledged
    - Changes in the text: AUC were converted to decimals
20. L188– The link to the app should be cited as a reference
    - Reply: The paper citation covers the citation of the web-app, which is an extension of the publication.
    - Changes in the text: N/A
21. L196 – It is not clear what this percentage refers to
    - Reply: The percentage refers to percent positive cores,
    - Changes in the text: N/A
22. L315 – The abbreviation has been already used before
    - Reply: Acknowledged
    - Changes in the text: The abbreviation is no longer spelled out
23. L326 – There's a space missing
    - Reply: Acknowledged
    - Changes in the text: The space was added
24. L329 – Whether the target volume planning objective was met or not due to a priority…
    - Reply: Acknowledged
    - Changes in the text: The phrase was changed to "Whether the target volume planning objective was met or not due to a priority"
25. L337 – Provide one or two examples of what these errors could be
    - Reply: Acknowledged
    - Changes in the text: "such as delivery of the desired dose"

Reviewer C

1. Categorize the methods for XAI in machine learning in a better way. Add a table with references and applications in oncology. e.g. machine learning method, what organ is it related to, how was the xai method utilized, the corresponding reference.
   - Reply: There are many way to organize the included studies could be organized. It is our opinion that the simplest way, and most beneficial way to a reader who is potentially most interested in a single discipline, is to organize it in the way we did. The review is of the benefits of explainability and how it has been applied in the literature. We have however added Table 3, which offers the suggested information, which allows the reader an alternative way to browse the included. We have also rearranged the sections to have more cohesive themes.
   - Changes in the text: Table 3 has been added and structure has been rearranged.
2. The authors only prove a survey of traditional machine learning classifiers. The are a lot explainability methods for deep neural networks. The authors only consider LIME and SHAP for machine learning. However, in Fig. 3 they are illustrating deep neural network's explanations.
   - Reply: There is a distinction between machine learning algorithms and machine learning explainability frameworks. The search criteria did not limit Figure 3 uses SHAP to explain a neural network. There are other model-specific frameworks for explainability. In the interest of limiting scope to broadly applicable machine learning explainability, the search query and papers discussed in this review are mostly limited to SHAP and LIME. We have updated the title and added clarification to the introduction that we focused on model-agnostic frameworks.
   - Changes in the text: N/A
3. Add better supporting figures. The quality of the figures currently seems low especially the SHAP analysis figures.
   - Reply: The figures were downloaded from the respective manuscripts so we are unable to improve their resolution. However, it is out opinion that they are of sufficient quality to read all components clearly.
   - Changes in the text: N/A
4. The authors need to do a critical account for SHAP and LIME frameworks for machine learning.
   - Reply: Please refer to the response to Reviewer A Point 2. Although we feel adding additional paragraphs to explain this is beyond the scope of this paper, we added a table summarizing the frameworks
   - Changes in the text: The core goal of both algorithms was added to the third paragraph of the introduction. We also added Table 1.

5. Is there a need of a framework for validating the explanations? Are they useful in the real life scenario? The authors need to discuss this. Sometimes, explanations lead to over diagnosis.
    - Reply: Given that explanations are merely representations of the model itself, as long as the model has been validated, so too are the explanations. An analogy would be Cox regression. We validate if the underlying model is appropriate. We do not validate the specific coefficients. At its core, explanations are conveying risk that need to be interpreted, so they do not inherently lead to overdiagnosis. Particular relationships identified by the ML model and highlighted via XAI can be further validated by "conventional" (i.e., univariate, parametric) statistical tests, and replicated in the separate datasets (when available).
    - Changes in the text: A sentence was added to the limitations section specifying that explanations are limited by the quality of the model itself.
6. The authors only provide figures related to SHAP analysis. They should include figures related to other explainability methods as well.
    - Reply: The fundamental concept behind figures generated by explainability methods is the same, regardless of method used. SHAP tends to be represented more commonly due to being newer and producing more attractive figures. The figures that were selected were primarily selected to give a diverse representation of applications of XAI. Further, we are limited in the publications we can use for figures due to fees, so had to either use figures from publications in open-access journals where we obtained permission or where an author on our review was also an author on the paper. For these reasons, we are leaving the selected figures as is.
    - Changes in the text: N/A

Reviewer D

1. In oncology, where patients' care, health, and beneficence matter, the reliability and accurateness of information and data quality are critical, which depends entirely on the data source. The close collaboration between Oncologists and computer scientists should be emphasized. There is a golden opportunity for electronic health records to be shared between hospitals or institutions to provide an integrative approach and reliable choice of care, maintaining data confidentiality and proper use of the information for oncology patients.
    - Reply: This is an excellent point worth adding to the future directions section.
    - Changes in the text: This point has been added to the future directions section.
2. This brings me to my next thought, which is an additional use of XAI with respect to prescriptive or semantic analytics; in other words, there is crosstalk amongst the

applications you so rightly described. The former can simulate outcomes for all possible scenarios; the latter recognizes the semantic relationships between data attributes to discover new information and possibly, use XAI to address complex interrogatives. The "Digital Twin" framework needs to be mentioned here as integrates individual-level data, such as proteome and clinical characteristics, with other factors like clinical trials and population studies to create multi-scale and multi-modal data sets for model training. Digital twins for predictive oncology will be a paradigm shift for precision cancer care and XAI can be perfect for model interpretability.

- Reply: Although this is an excellent point, we found it challenging to fully discuss the digital twin framework without getting too off topic. Therefore we have briefly added to our future directions section that this also has implications for precision medicine due to massive amounts of patient data made available by the EHR.
- Changes in the text: We have added a sentence regarding potential avenues to implementing XAI with precision medicine.

3. Is it worth mentioning the potential economic benefits, in an era of stringent economics; that is, by implementing AI forecasting models and XAI frameworks, they can explain how different policy scenarios can be explored, and favor the most effective strategy to enhance Oncology economics?
    - Reply: Although this is an excellent point regarding the potential of ML in general, which explainability no doubt would help, a nuanced discussion of this topic would deviate from the scope of this review. However, we do feel it is reasonable to include in a list of the many things ML+explainability might improve in the modern information era.
    - Changes in the text: We have added "more economic" to the list of ways ML can improve oncology in the future directions section.
4. Ultimately, there will be some challenges as XAI developers will realize that XAI systems are affected by human interactions and human-like traits; hence, empirically backed principles should be uniformly applied for their safer and potentially more effective design.
    - Reply: This is a good point to include when describing appropriate models.
    - Changes in the text: The importance of explaining empirically sound methods has been added to the limitations section.

Reviewer E

1. The paper is lacking a comprehensive introduction/explanation of how the two frameworks for XAI work. I imagine it will be very hard to realise the potential of this

technology without understanding how it work, how it can be integrated with machine learning models and what type of outputs it can produce.

- Reply: An overarching explanation of how the explainability frameworks work is beyond the scope of this paper, as it is a nuanced and complicated topic that would take a whole separate paper to describe. This is accomplished in the citations by Ribeiro et al and Lundberg et al. The key point for the reader needs to understand is described in the second paragraph of the introduction. However we added table 1 to summarize the two included frameworks
- Changes in the text: We have added a sentence to the second paragraph of the introduction to provide further clarification of the key concepts of XAI. We also added Table 1 to summarize the frameworks.

2. The title of the paper is: "Applications of explainable artificial intelligence frameworks in oncology…" which suggests that that the paper talks about how XAI can be used in oncology and what the benefits are. Instead the paper list all existing works that used one of the two frameworks, mainly for the same purpose i.e. to identify the key features/characteristics relevant to a specific task. There was a lot of repetitive content, where the same finding was reported (XAI can be used for the analysis of feature importance and feature selection) just with different studies. As it stands, the paper provides an overview of medical findings that were delivered with application of XAI technology. If this was the intention of the authors the title of the paper should be changed.

- Reply: It is our intention to provide examples of how XAI has been applied in the literature. We agree that this can be repetitive, given that it is similar plots and core concepts that are being described. We have changed "Applications" to "Utilization" since what the reviewer describes was our intention. The new Table 3 also helps emphasize how XAI was used.
- Changes in the text: Applications" was changed to "Utilization" in Oncology. Table 3 was added.

3. Instead of going through different fields in oncology and listing all the papers that used the XAI frameworks, it would make more sense to have different sections talking about different ways the XAI frameworks have been applied in oncology (i.e., for feature selection, identification of prognostic thresholds etc.).

- Reply: We agree that there are other ways that the paper could be organized. It is our preference to stay with our present structure for two reasons. First, readers may want to explore applications of XAI by field and not specific application. Second, in many papers the actual application might be multi-fold, such as both helping select thresholds, helping select relevant features, and inspiring confidence in the model by illustrating how it makes its predictions. Classifying many studies would get muddy. The suggested scheme would be more useful for the CS-type audience, whereas our original scheme is a better fit with the oncology audience (which is our primary "target"). For these reasons, we are not presently restructuring the paper in that format. However, we have added sentences to the abstract and conclusions on how XAI can be harnessed to do what this reviewer described. We also summarized the studies in Table 3. Lastly we did restructure the respective sections to have better thematic threads regarding specific uses of XAI

- Changes in the text: Table 3 was added. Abstract and conclusions were updated.

Reviewer F

1. XAI techniques are modality specific (ie structured data, image data, series data, natural language data). In structured data analysis, XAI aims to identify the variables which influenced the model output. In image analysis, XAI aims to identify the regions of interest which influenced the model output. The XAI techniques discussed in this study apply primarily to structured data analysis and if the scope is limited to structured data analysis, this should be specified in the title.
   - Reply: The search was not limited to structured data, though that does comprise a large proportion of such studies due to availability of structured datasets and difficulty in creating large imaging datasets. However, several included studies included image analysis including the study highlighted in Figure 3 and studies in the imaging and pathology sections. However, given that the query only included "SHAP" and "LIME", and the emphasis of the paper is on model agnostic frameworks, we updated the title, clarified the scope in the into, and removed studies that don't use SHAP or LIME
   - Changes in the text: we updated the title, clarified the scope in the into, and removed studies that don't use SHAP or LIME
2. Abstract Line 33. Please reword "XAI can break transform complicated ML models into easily understandable charts and interpretable sets of rules". Current XAI methods provide a simplified representation of model reasoning, which is not guaranteed to align exactly with the true decision functions.
   - Reply: The two statements are not incompatible. XAI does not just "simplify", but also (and more often than not) brings into relief the most salient components of the complex model. This, of course, does not have to compromise the original model's performance. The statement in the abstract is still valid. Further, for certain ML models, SHAP can exactly explain the

model (https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/explainers/Exact.html)

- Changes in the text: N/A

3. It should be specified that XAI techniques are modality specific (ie structured data, image data, series data, natural language data). In structured data analysis, XAI aims to identify the variables which influenced the model output. In image analysis, XAI aims to identify the regions of interest which influenced the model output.
    - Reply: This is a great point
    - Changes in the text: A sentence was added to the introduction with these details.
4. For every XAI method mentioned in the diagram, a brief, plain-english explanation of the methodology should be provided. This would ideally be complemented by a diagram to aid understanding. Aim to teach readers who are unfamiliar with XAI to have some conceptual insight into how XAI results are generated.
    - Reply: It is not clear which diagram this refers to. A general overview of how XAI is intended to work has been added to the in introduction. As detailed by responses to other reviewers, a more specific explanation of how the specific XAI frameworks work in the text is beyond the scope of this review, but is available in the included relevant citations as well as the newly added Table 1.
    - Changes in the text: The introduction has been supplemented to give a brief overview of the general principle of how XAI is meant to work. Table 1 has been added.
5. Please delete the term “comprehensive”, which must be substantiated by full systematic review results (including screening) on all relevant databases (scopus, embase, sciencedirect).
    - Reply: Acknowledged
    - Changes in the text: Comprehensive has been deleted
6. Please provide the earliest search date, or specify that studies were included from the beginning of the database.
    - Reply: Acknowledged
    - Changes in the text: “Beginning of database” was added to Table 1.
7. Please provide the exact search query.
    - Reply: The exact search query is included in Table 1
    - Changes in the text: N/A
8. Please provide some of the quantitative XAI results (ie authors’ visually inspected the SHAP interaction plot, identifying a “U-shaped” distribution between gleason and PSA at low PSA scores, indicating that the interaction was nonlinear in this

subcohort) and quantitative XAI results (when included in a cox regression model PPC >= 50% interacted significantly with Gleason score >=8 (P<.001)).

- Reply: Where relevant, available quantitative metrics were included (AUC for example). Example: "Their model using lymph node ratio outperformed the AJCC schema at predicting 1 year overall survival with an area under the curve (AUC) of 0.638 versus 0.586." There are no true quantitative XAI results since it is a qualitative modality. ML models do not yield summary statistics like Cox regression does. So the statistics requested are only possible if the type of validation is subsequent Cox (or comparable) regression, which is not common practice. This is further detailed in the reponse to point 10.
- Changes in the text: N/A

9. Please specify the value of applying XAI in these studies (ie identifying nonlinear interactions, which may not have been detectable using standard bivariate measures such as the correlations, or multivariate measures such as the partial correlation matrix)
   - Reply: The discussions have sentences on the key takeaway rendered by XAI. AN example from the paragraph discussing Li et al: "modeling with XAI revealed nuances that contradict modern risk stratification. Visualization of such interactions is only possible by using a more complicated model than standard regressions and then explaining that model."
   - Changes in the text: N/A
10. Please provide some critique of the included XAI studies. Discuss whether articles' XAI results are reliable or unreliable:
    i) Were quantitative results generated, or were results based purely on inspection of plots?
    ii) Were significance tests performed, or are they available? In most cases permutation tests would be applicable.
    iii) Have you any reason to believe that the results were unreliable – ie small sample sizes, highly imbalanced data or systematic bias within the subgroups examined?
    iv) Were XAI results validated internally or in data from separate institutions?
    v) Was this the most appropriate XAI method to apply in this case and why?
    - Reply: Please refer to the response to Reviewer C Point 5. XAI simply provides an explanation of the underlying model and do not have reliability or significance tests. The attributes this point is getting at are the attributes of the machine learning model themselves. These can be evaluated with any number of tests including AUC, precision, recall, sensitivity, specificity, etc. These are performed by validating the model in a holdout testing dataset that was not used to train the model. The studies almost all report such relevant metrics.

Further, XAI is inherently a qualitative entity that can guide subsequent quantitative analyses not included under the XAI umbrella, such as confirmatory Cox regression. Lastly, since at their core the XAI frameworks accomplish the same goal, one is not inherently more appropriate than another. It is a matter of preference.

- Changes in the text: N/A

11. Line 124. I suggest deleting any line in the format "Another such example of where XAI is beneficial is…". XAI is generally applicable, these statements are unnecessary.
    - Reply: Acknowledged
    - Changes in the text: The introductory sentence specified was removed.
12. Line 138. Please provide significance results (or lack thereof) if a claim of one model outperforming another is made. Where AUC results are compared a DeLong test is a minimum requirement. The original study omitted this.
    - Reply: Acknowledged
    - Changes in the text: As stated, the study did not perform DeLong's test or include confidence intervals. Therefore a qualifier that such test was nit performed was added.
13. Lines 166-180. This is an extremely brief mention of one of the most frequent applications of XAI – CNN image analysis. Mentioning two results without any explanation of methods etc is insufficient for such an important field, see the following reviews (https://doi.org/10.1016/j.media.2022.102470, https://doi.org/10.1016/j.inffus.2021.07.016, DOI:10.1016/j.compbiomed.2021.105111). The most commonly applied XAI frameworks in clinical imaging studies according to a recent systematic review were CAM and GRAD-CAM ("Explainable artificial intelligence (XAI) in deep learning-based medical image analysis", Van der Velden 2020, https://doi.org/10.1016/j.media.2022.102470). Neither of these methods are mentioned in this paper, which raises a concern regarding how comprehensive this review is.
    - Reply: Given these papers used specific frameworks not previously detailed in the review, we specified limited focus to model-agnostic frameworks in the title and introduction and removed these references.
    - Changes in the text: The introduction was updated and these studies were removed.
14. I would advise to omit the CNN image analysis section from the paper and acknowledge limitation of the scope to XAI for structured data analysis.
    - Reply: Although it would be reasonable to omit the section on studies that used CAM and Grad-CAM as detailed in point 13, several other included CNN were analyzed using SHAP and LIME, which are more heavily

represented in this paper. Therefore, for the sake of completeness, it is our preference to not broadly remove CNN. We did however remove studies not dealing with SHAP or LIME.

- Changes in the text: Studies were removed.

15. Lines 210-222. Some critical appraisal must be performed to decide which studies to include, please see the radiomics quality score (RQS) (https://www.nature.com/articles/nrclinonc.2017.141/tables/1) or the TRIPOD checklist (https://www.equator-network.org/reporting-guidelines/tripod-statement/). Although the Manikis study achieves a reasonable 14/36 RQS points, Kha Q-H study would score considerable fewer. The KHA-QH study also applies feature selection with the Pearson correlation coefficient outside of the cross validation, a well known cause of overfitting in radiomics studies (doi: 10.1186/s13244-021-01115-1). Thus, presenting the features selected in a single radiomics study, in a field where feature selection is so variable (doi: 10.1186/s13244-022-01245-0), is both inappropriate and irrelevant to the discussion of XAI.
    - Reply: The focus of this study is not a critical appraisal of radiomics studies. It is focused on showing how XAI is used in published literature. So assessing these studies with RQS is beyond the scope of this review. However, we agree listing the features is not needed.
    - Changes in the text: We removed the relevant features from the text.
16. Lines 258-261. Please revise any line with the format “using XAI, features xyz were identified as predictors”. The overall results of individual studies are of minor relevance to this paper, it is much more important to discuss how XAI methods were applied, and the XAI-specific results which supported the overall study findings (ie in this study describe the B cell, T cell cd8+, macrophage and T cell NK SHAP plots and the characteristics which suggested feature importance).
    - Reply: Acknowledged. The interpretation has been rephrased to change emphasis.
    - Changes in the text: The text has been revised to better highlight interpretation of the plots.
17. Lines 378-393. This is a good summary and I agree with the recommendations here overall, though please note that the “simple” models should include regression models such as LASSO.
    - Reply: This is a good point. Indeed “simple” models such as LASSO are likely underutilized and may yield favorable results without needing to turn to more complicated and less interpretable ML models.
    - Changes in the text: Regression models such as LASSO was added to this sentence.

18. Lines 397-399. This is an important topic which is highly relevant to the paper. However, it is only briefly mentioned here. Please discuss this topic in more detail, ie studies which apply lasso or Bayesian networks.
    - Reply: Although this is relevant to the paper, it is our opinion that it is discussed in appropriate depth for the scope of the paper. The topic of the review is specific to XAI frameworks (model-agnostic specifically), which are distinct from models designed to be inherently interpretable.
    - Changes in the text: N/A
19. A narrative review of should leave the reader with some understanding of XAI – how it works, different approaches, opportunities it creates (ie more reliable and interpretable decisions, model validation, discovery of new biology) and the risks it presents (misrepresentation of the model's true reasoning). Listing conclusions of XAI application in clinical research, confirms that this is a valuable field of research, but provides little knowledge which the reader can apply (Imagine teaching medical students about breast cancer by telling them that a breast cancer cohort died earlier than a group of healthy controls, they need to know how cancer develops, what risks arise, how to assess them, how to treat them. Likewise, with XAI, clinicians need to know how to pick a method for their study, how to interpret results, how to know if the XAI results are reliable etc).
    - Reply: The intent behind the structure of this review was to provide a comprehensive overview of examples of how such analyses are done in the literature, and then put them in context. The manuscript has been revised to further facilitate this, including grouping studies by general XAI intent. The challenge is this review inherently describes a visual and qualitative discipline in text, so will require the reader to look at some cited papers to get a more complete understanding.
    - Changes in the text: The text, future directions and limitations have been updated to help with the utility of the narrative review.
20. Consider whether XAI is superior to traditional feature selection methods such as LASSO? Have studies compared these? What are the advantages conveyed? Are these worth the additional overfitting risk posed by XAI models?
    - Reply: This is also relevant and worth adding to the limitations section.
    - Changes in the text: We added the importance of model selection in the limitations section.