Peer review file

Article information: https://dx.doi.org/10.21037/jtd-21-1107

Reviewer A

Major comments:

Comment 1: All tables and figures need a more detailed explanation and caption, including a description of the shortcuts/terms used in the tables/figures, so that the non-expert reader can easily understand what the authors are referring to. For instance:

Reply 1: Thanks so much for this insightful recommendation. We have modified and explained each point as follows.

Point 1: In Figure 1 the authors should define what the lambda_J1 and lambda_j2 actually means. In addition, unless I am missing something, the information about the number of trials is missing.

In the text (line 166) they mention that "in the first several trials, optuna randomly sampled the search area, then gradually concentrated on the areas with smaller objective value (i.e., logistic loss in this paper), and finally found the best ones". However, it is not clear what you refer to by "trials", is it the number of iterations for a specific validation fold? Where is the information about the "trials" (or iterations) in the figure?

The value of the objective function is represented by the different blue colors, but still the value of the objective function does not always correlate with increasing number of iterations since there can be fluctuations during the optimization. A way to display this could be to use different markers for different ranges of iterations, e.g. triangles for iterations 0-5, circles for iterations 5-10, etc.

Reply 1.1: As suggested, we add the parameter descriptions of the training parameters presented in Table 2, as displayed in the Table 4 in the Supplements Appendix (see page 29, line 470), where lambda_l1 and lambda_l2 are included.

Additionally, the number of trials, 100, has been referred to in page 9, line 180. To highlight this information, we add it to the legend of Figure 1 (see page 13, line 205).

In the revised manuscript, we clarify the definition of trial (see page 9, line 174 to 179). A trial is an iteration in the tuning procedure. And the number of trials is set to be 100, so there are 100 iterations in a hyperparameter tuning process.

Moreover, thank you for pointing out our mistake. It is truth that the value of the

objective function does not always correlate with increasing number of iterations since there can be fluctuations during the optimization. The description in the sentence “in the first several trials…” is not precise. To better illustrate Figure 1, we rephrase the section, Results - Hyperparameter tuning, as shown in page 15, line 210 to 224.

Figure 1 only shows the sampled dots in the parameter search space, i.e., the selected hyperparameters in each iteration of the hyperparameter tuning process. Figure 1 is generated automatically, and we have not recorded the iterative process corresponding to Figure 1. Also, each execution of the hyperparameter tuning procedure would provide a different iterative process and result. So, it is hard to know which iteration does the dot correspond to.

Changes in the text:

Page 29, line 470: Table 4: The parameter descriptions of the training parameters of XGBoost, CatBoost, LightGBM, GBDT and RF

Page 13, line 205: “There are 100 dots in the picture.”

Page 9, line 174 to 179: “A trial is a single execution of the objective function which is defined as the average logistic loss of the 5-fold cross-validation of the model in this paper. In each trial, the hyperparameters are selected from the parameter space according to the prior information, and then the stratified 5-fold cross-validation is executed to produce the average logistic loss to estimate the model with the selected hyperparameters.”

Page 15, line 210 to 224: “Figure 1 provides a diagram of the process hyperparameter tuning of LightGBM model. Lambda_l1 and lambda_l2 are two hyperparameters in LightGBM model. There are 100 dots in Figure one and each dot represents a trial, whose location shows the corresponding lambda_l1 and lambda_l2 values. The shade of blue indicates the range of the objective values for the trials. In addition, it can be observed that the lighter the color, the denser the dots. Because Bayesian optimization balances between exploration (hyperparameter configuration for which the objective value is most uncertain) and exploitation (hyperparameter configuration expected close to the optimum). In other words, some of the trials might concentrate on hyperparameter values around the local minimum, the others would try new hyperparameter configurations. Therefore, in the area with low objective value, the dots would assemble, and the hyperparameters near the dots with high value would not be selected to have a trial.

The best hyperparameters found in the hyperparameter tuning processes of XGBoost, CatBoost, LightGBM, GBDT and RF are shown in Table 2. The descriptions of all training parameters are displayed in Table 4 in the Supplements Appendix.”

Point 2: In Table 1, it would be good to refer to a dictionary for SEER terms, if that exists. Something similar to this link:

https://staging.seer.cancer.gov/cs/schema/02.05.50/esophagus/?breadcrumbs=(~schema_list~), another option is this document: https://seer.cancer.gov/data-software/documentation/seerstat/nov2017/TextD ata.FileDescription.pdf

Reply 1.2: As suggested, we have modified the feature names in Table 1 (see page 10, line 195) referring to a dictionary for SEER terms. Besides, the feature names in Figure 3,4,5 (see line 245, 250 and 284 respectively) are also modified.

Changes in the text:

Table 1: Selected clinicopathologic features from SEER dataset. (See page 10, line 195)

Figure 3: SHAP feature importance measured as the mean absolute SHAP value. (See page 18, line 245)

Figure 4: SHAP summary plot. Each point is a Shapley value for a feature and an instance. (See page 19, line 250)

Point 3: In Table 2, the parameters should be defined so that the non-expert / clinical readers can better understand what you are referring to. If it is too long, and the journal allows for it, maybe you can define all parameters in a supplementary material.

Reply 1.3: As suggested, we add the parameter descriptions of the training parameters presented in Table 2, as displayed in the Table 4 in the Supplements Appendix (see page 29, line 470).

Changes in the text: Table 4: The parameter descriptions of the training parameters of XGBoost, CatBoost, LightGBM, GBDT and RF. (See page 29, line 470).

Point 4: Figure 5 needs clarification (maybe a legend) about what the histogram in light grey represents and what the blue points represents. I assume one the blue points represent the SHAP and the light grey the values of the features but it should be clarified in a legend.

Also in the text they mention “except the values with special meaning” → what do you mean with special meaning? Please clarify.

Also in Figure 5.b you also need some SHAP values in between 200 and 1000, do you consider these outliers? How do you explain this?

Reply 1.4: The clarification of the blue dots and grey histograms has been added to the legend of Figure 5 (See page 21, line 284).

In addition, the two tables below show the description of the special value, 95-99 in Regional nodes positive and 990-999 in CS tumor size. Also, we have briefly generalized the two tables and explained “except the values with special meaning” in page 20, from line 276 to 282.

REGIONAL NODES EXAMINED (1988+)

Field Description: Records the total number of regional lymph nodes that were removed and examined by the pathologist.

Code	Description
00	No nodes were examined
01-89	Exact number of nodes examined
90	90 or more nodes were examined
95	No regional nodes were removed, but aspiration of regional nodes was performed
96	Regional lymph node removal was documented as a sampling, and the number of nodes is unknown/not stated
97	Regional lymph node removal was documented as a dissection, and the number of nodes is unknown/not stated
98	Regional lymph nodes were surgically removed, but the number of lymph nodes is unknown/not stated and not documented as a sampling or dissection; nodes were examined, but the number is unknown
99	Unknown whether nodes were examined; not applicable or negative; not stated in patient record
126	Blank

CS TUMOR SIZE (2004-2015)

Field Description: Information on tumor size. Available for 2004-2015 diagnosis years. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS.

Code	Description
000	Indicates no mass or no tumor found; for example, when a tumor of a stated primary site is not found, but the tumor has metastasized.
001-988	Exact size in millimeters
989	989 millimeters or larger
990	Microscopic focus or foci only; no size of focus is given
991	Described as less than 1 cm
992	Described as less than 2 cm
993	Described as less than 3 cm
994	Described as less than 4 cm
995	Described as less than 5 cm
996-998	Site-specific codes where needed
999	Unknown; size not stated; not stated in patient record
888	Not applicable
1022	Blank

Besides, in Figure 5.b, for the dots with CS tumor size values in between 200 and 1000, the number of these outliers is small compared to the total sample size, so it is hard to exactly explain their SHAP values. Therefore, these outliers are considered with the dots with CS tumor size values in between 100-200 together.

Changes in the text:

Page 21, line 284: Figure 5: SHAP dependence plot, (a) for Regional nodes positive

(1988+), (b) for CS tumor size (2004-2015). This picture plots the SHAP value of the feature vs. the value of the feature for all the patients in the dataset. The light grey bars are the frequency distribution histograms for the two features.

Page 20, line 276 to 282: "Value 95-99 in Regional nodes positive refer to the case that no regional nodes were removed or the case that the number of nodes is unknown/not stated. Value 991-995 in CS tumor size refer to the cases that tumor size was described as less than 1 cm to 5 cm respectively. Value 990 means microscopic focus or foci only; no size of focus is given, value 996-998 mean that site-specific codes where needed, and value 999 means unknown. The figures displayed that except the special values mentioned above, the higher values the features the higher risk of death."

Comment 2: About the methodology, a major comment I have is about the training of the models. The authors mention that they do a 5-fold cross-validation, but many details are missing and need to be clarified.

Q1: How many patients did you use for training and test in each fold?

Q2: Did you do the cross-validation for parameter tuning and then have another independent test set to generate the final results?

Q3: The text does not mention a text set anywhere, which is important to build and deploy a model correctly.

Q4: In case you do have a test set, did you train the models with all the datasets with the selected parameters after cross-validation, or did you use one of the validation folds (e.g. the best) as your model?

Reply 2: Thanks so much for this insightful recommendation. We have rephrased the section, Methods - Model evaluation, which can response Q1 and Q3 (see page 8, line 143 to 152).

In the process of parameter tuning, a 5-fold cross-validation is executed in each trial to produce the average performance measure result to estimate the model with the selected hyperparameters. After hyperparameter tuning, we get the best hyperparameters. And then, a 5-fold cross-validation is executed to produce the average performance measure result for each compared model in the final model comparison. We have added some explanation (see page 9, line 176 to 182) to respond Q2, and responded Q4 in page 17, line 234 to 236.

Changes in the text:

Page 8, line 143 to 152: "K-fold cross-validation is the most commonly resampling techniques used in evaluating ML models. The original sample is randomly divided into k equal sized subsamples. Among the k subsamples, a single subsample is held as the

validation data to test the model, and the residual k − 1 subsamples are used as training data. The cross-validation process is repeated k times, with each of the k subsamples used exactly once as the test data. The k results can then be averaged to produce a single estimation, i.e., the performance measure of the model. In this paper, for the dataset with unequal class proportions, stratified k-fold is used, where the folds are made by preserving the percentage of samples for each class."

Page 9, line 176 to 182: "In each trial, the hyperparameters are selected from the parameter space according to the prior information, and then the stratified 5-fold cross-validation is executed to produce the average logistic loss to estimate the model with the selected hyperparameters. The parameter spaces of each models are shown in Table 2 and the number of trials is set to be 100. After 100 trials, the hyperparameters with minimum average logistic loss is chosen for the final model comparison."

Page 17, line 234 to 236: "In Table 3 and Figure 2, we summarized that the performance of eight models in terms of ROC_AUC, accuracy, logistic loss and precision-recall curve, which are average results of 5-fold cross-validation."

Comment 3: The authors mention in the introduction and discussion that there is a previous work (very well known, Sato et al. 2005) that used ANN for prognosis prediction of esophageal cancer and achieved a better accuracy (ROC_AUC: 0.88). The authors associate the slightly better accuracy to the increased number of features used in this work (Sato et al. 2005), which makes sense. However, the model might also play an important role. It would be very interesting to see the comparison of an ANN (trained with the data used in the present study) with the presented methods. Do you think an ANN would improve your results? Apart from the simplicity/interpretability, what is the added value of using classical machine learning methods instead of deep learning (e.g. ANN)? The advantage of using ANN is that the feature selection process would be directly embedded into the model training.

Reply 3: Thanks so much for this insightful recommendation.

As suggested, we use ANN model to predict the 5-year survival status of patients for comparison. The ANN model used is chosen from 14 different ANN structures (n-2-1, n-3-1, n-4-1, n-5-1, n-6-1, n-2-2-1, n-2-4-1,n-2-6-1, n-4-2-1, n-4-4-1, n-4-6-1, n-6-2-1, n-6-4-1, n-6-6-1; n = the number of features) as in the study of Sato et al. (2005) (see page 7, line 132 to 135). The experiment results show that n-4-4-1 performs best, as shown in Table 1. We think Table 1 is not important for this paper, so only the result of n-4-4-1 structure have shown in Table 3 (see page 16, line 233) in the revised manuscript.

As displayed in Page 16, line 226 to 233 in the revised manuscript, the performance of ANN is slightly worse than that of XGBoost, LightGBM, CatBoost and GBDT. In addition, XGBoost, LightGBM, CatBoost, GBDT and RF are all belong to Ensemble

Learning. This paper is more focus on Ensemble Learning, which is recognized best-in-class in the field of machine learning when it comes to small-to-medium structured/tabular dataset. However, with the development of deep learning technique in recent years, I believe that new technique in deep learning field worth trying, which will be our research focus in the future.

Table 1: Model performance using 14 different ANN structures.

	n-2-1	n-3-1	n-4-1	n-5-1	n-6-1	n-2-2-1	n-2-4-1	n-2-6-1	n-4-4-1	n-4-6-1	n-6-2-1	n-6-4-1	n-6-6-1
AUC	0.500	0.501	0.837	0.836	0.839	0.801	0.832	0.834	**0.844**	0.838	0.841	0.839	0.840
accuracy	0.855	0.855	0.869	0.868	0.868	0.869	0.868	0.869	**0.871**	0.869	0.871	0.872	0.871
logistic loss	0.415	0.414	0.312	0.313	0.311	0.324	0.319	0.315	**0.308**	0.312	0.310	0.311	0.312

Changes in the text:

Page 7, line 132 to 135: "Specially, the ANN model used in this paper was chosen from 14 different ANN structures (n-2-1, n-3-1, n-4-1, n-5-1, n-6-1, n-2-2-1, n-2-4-1,n-2-6-1, n-4-2-1, n-4-4-1, n-4-6-1, n-6-2-1, n-6-4-1, n-6-6-1; n = the number of features) as in the study of Sato F (4)"

Page 16, line 226:

Figure 2: Visual presentation of model performance based on eight algorithms trained by the complete dataset. (a) displays the precision-recall curve. (b) displays the ROC curve. When the area under curve is closer to 1, the performance of model classification and prediction is better.

Page 16, line 233: Table 3: Model performance using eight algorithms.

Page 21, line 295 to 300: "In this study, discrimination of the performance of the three newly developed variants of GBM was similar, with little variability in ROC, PRC (Figure 2) and other three metrics (Table 3). They outperform the other models, including the ANN used by Sato F (4). The predicting result (ROC_AUC=0.88) in the study of Sato F (4) is higher than that in this paper, but there are more features (199 features) used to train the model providing more information thus better predicting outcomes."

Comment 4: Regarding the hyperparameter optimization, it is mentioned that a Bayesian optimization is used and referred to a couple of papers. However, I think that more information is needed in this part, for instance, about the prior distributions assumed for the features in the Bayesian framework.

Reply 4: We deeply appreciate for this point. We have rephrased the section, Methods

- Hyperparameter tuning, to clarify the concept of Bayesian optimization (see page 8, line 157 to 182). We first introduced two traditional ways of performing hyperparameter optimization, grid search and random search for comparison, which do not learn from the previous trial results. We then explained how Bayesian optimization works. Moreover, in the previous version of the manuscript, we have not clarified the meaning of prior information. The prior information here refers to the hyperparameter configurations and the corresponding objective function loss of the model in the all prior trials.

Changes in the text:

Page 8, line 157 to 170: "The two traditional ways of performing hyperparameter optimization are grid search and random search. Grid search performs an exhaustive search through a manually specified subset of the hyperparameter space, which is computationally very expensive. In random search, the hyperparameters are randomly selected not every combination of parameter is tried. These two methods do not learn from previous results. Conversely, Bayesian optimization iteratively evaluates a promising hyperparameter configuration based on prior information, including previous hyperparameter configurations and the corresponding objective function loss of the model, and then updates it. Bayesian optimization allows exploration (trying new hyperparameter values) and exploitation (using hyperparameter configuration resulting in the lowest objective function loss) to be naturally balanced during the search. In practice, it is shown that compared to grid search and random search, Bayesian optimization is able to obtain better results in fewer evaluations, due to the ability to reason about the quality of trials before they are run."

Page 9, line 172 to 182: "Using Optuna, we can define the parameter space and the trials by Python syntax, and adopt state-of-the-art algorithms for sampling hyperparameters and efficiently pruning unpromising trials. A trial is a single execution of the objective function which is defined as the average logistic loss of the 5-fold cross-validation of the model in this paper. In each trial, the hyperparameters are selected from the parameter space according to the prior information, and then the stratified 5-fold cross-validation is executed to produce the average logistic loss to estimate the model with the selected hyperparameters. The parameter spaces of each models are shown in Table 2 and the number of trials is set to be 100. After 100 trials, the hyperparameters with minimum average logistic loss is the best one, and are used in the final model comparison. Note that there are fewer hyperparameters used for SVM and NB, so obtained by trial-and-error method."

Comment 5: Data imbalance → The authors are aware that the database is unbalanced: "Besides, the Precision-Recall curve shows that the three variants are effective in predicting the imbalanced dataset, however, NB and SVM have poor performance in the class accuracy of small number of samples caused by the imbalance of sample

number."→ Please specify what type of imbalance you refer to ("dead" versus "alive"? Imbalance in the feature values?). In addition, how is this imbalance represented in the training, validation and testing subdatasets?

Reply 5: Thanks so much for the valuable comment. The type of imbalance we refer to in this paper is the imbalance of "dead" and "alive". We have modified the sentence in page 13, line 197 to 199 to highlight it. Additionally, we have explained the **stratified k-folds cross-validation** we used (see page 8, line 150 to 152), which present how this imbalance represented in the training, validation and testing subdatasets. Compared with the holdout cross-validation, where the dataset is divided into training set, validation set and testing set, the k-fold cross-validation is much less sensitive to the split of the training set, validation set and testing set.

Changes in the text:

Page 13, line 197 to 199: "The sample consists of two classes (9048 cases with "Dead" status and 1540 cases with "Alive" status), which showed the imbalance of sample number."

Page 8, line 150 to 152: "In this paper, for the dataset with unequal class proportions, **stratified k-folds** is used, where the folds are made by preserving the percentage of samples for each class."

Discussion:

Comment 6: Q1: The authors mention that "SEER Program data provided may not be able to the characteristics of each patient accurate expression (such as a large amount of lack of characteristic variables, complication, the specific mode of operation, postoperative complications of case did not reflect, etc.), and crucially, radiation and chemotherapy in patients with information (including whether the radiation and chemotherapy as well as the concrete plan of radiation and chemotherapy) is missing." → This paragraph is very badly written and should be rephrased.

Q2: In addition, it would be good to add examples and references of the features/scenarios they mention. For instance, cardiac and pulmonary complications from radiotherapy treatments: https://pubmed.ncbi.nlm.nih.gov/25554226/, https://pubmed.ncbi.nlm.nih. gov/31630867/

Q3: Please, discuss also how your model could be adapted to take into account these features in case they are available. Maybe a transfer learning strategy? Or training the model from scratch by adding more features?

Q4: What if some extra information is in the form of images (like computed tomography)?

Reply 6: Thanks so much for the valuable comment. For Q1 and Q2, we have rephrased the badly written sentence and add examples and references of the features/scenarios (see page 24, line 350 to 364, and page 29, line 456 to 468).

For Q3, we have added the discussion how our model could be adapted to take into account these features in case they are available in page 24, line 367 to 369.

Moreover, for Q4, we know very little about image recognition, so we are incapable of expanding on this in the manuscript. We suspect that the information in the form of images should be tackled by feature extraction and feature selection technique. Next, the image information will be considered with the tabular data together.

Changes in the text:

Page 24, line 350 to 364: "Some critical factors that are strong predictors of survival and patient outcome are unavailable in the SEER database, such as methods of surgery, postoperative complications, and more importantly, radiation and chemotherapy information. Some technical advances in surgical methods such as preoperative simulation, robot-assisted thoracoscopic esophagectomy and intraoperative real-time navigation may decrease the morbidity and mortality rate of surgery for esophageal cancer and hopefully improve oncological outcomes (28). Prone to a variety of complications is also one of the characteristics of esophageal cancer. Anastomotic leaks, chyle leaks, cardiopulmonary complications, and later functional issues after esophagectomy may result in long-term sequelae and even death (29). The effect of radiotherapy and chemotherapy on patients with esophageal cancer is still a research hotspot. For locally advanced esophagogastric junction patients, neoadjuvant chemoradiotherapy has better survival rate than neoadjuvant chemotherapy (30). At the same time, radiotherapy or neoadjuvant chemoradiotherapy may also increase the incidence of cardiac and pulmonary complications (31,32). These factors have a negative impact on the accuracy of the prediction."

Page 29, line 456 to 468:

"28. Beukema JC, van Luijk P, Widder J, et al. Is cardiac toxicity a relevant issue in the radiation treatment of esophageal cancer? Radiother Oncol. 2015 Jan;114(1):85-90.
29. Igor, WM, Sushanth, R, Anne OL. Complications After Esophagectomy. Surgical Clinics of North America. 2019.
30. Thomas M, Defraene G, Lambrecht M, et al. NTCP model for postoperative complications and one-year mortality after trimodality treatment in oesophageal cancer. Radiother Oncol 2019;141:33-40.
31. Kikuchi H, Takeuchi H. Future Perspectives of Surgery for Esophageal Cancer. Ann Thorac Cardiovasc Surg 2018;24(5):219-222.
32. Li J, Zhao Q, Ge X, et al. Neoadjuvant chemoradiotherapy improves survival in locally advanced adenocarcinoma of esophagogastric junction compared with

neoadjuvant chemotherapy: a propensity score matching analysis. BMC Surg 2021;21(1):137."

Page 24, line 367 to 369.: "If the features mentioned above is available, the model should be trained from scratch, including the process of feature selection, hyperparameter tuning, and model evaluation."

Comment 7: The authors used SHAP as a model to interpret the feature importance, but there are many other models that might be used too, such as LIME. Discuss, if any, the limitations of SHAP and if any other state-of-the-art model could be used to retrieve the feature importance.

Reply 7: We sincerely appreciate the excellent suggestion. In the revised discussion, we briefly discuss a few pros and cons of SHAP and LIME, and explain the reason why we choose SHAP (see page 22, line 315 to 324).

Changes in the text: (page 22, line 315 to 324) "SHAP and LIME (Local Interpretable Model-agnostic Explanations) are both popular approaches for model explainability. LIME builds sparse linear models around each prediction to explain how the black box model works in that local vicinity. In the NIPS paper (21), the authors of SHAP show that SHAP provides the only guarantee of accuracy and consistency, while LIME is actually a subset of SHAP but lacks the same properties. However, SHAP is an exhaustive method considering all possible predictions for an instance using all possible combinations of inputs, and thus time consuming compared with LIME. But the sample size 10588 is a small dataset in data mining fields, thus SHAP is used in this paper."

Minor comments:

- Line 153 → We compared

- Line 219 → Discussion

- Line 268 → the specific mode of operation, did you mean the specific treatment applied? Please clarify

Reply 8: We were really sorry for our careless mistakes. Thank you for your reminder. We have made the corrections in our resubmitted manuscript. The specific mode of operation is revised as methods of surgery.

Changes in the text:

Page 13, line 201: "We compared the performances of each models trained by the complete dataset and the dataset removed the non-significant features."

Page 21, line 290: "Discussion"

Page 24, line 350 to 353: "Some critical factors that are strong predictors of survival and patient outcome are unavailable in the SEER database, such as methods of surgery , postoperative complications, and more importantly, radiation and chemotherapy information."

Reviewer B

Major comments:

Comment 1: Why the final XGBoost model has not been named differently to avoid confusion?

Reply 1: Thank you for your valuable comments on our article. The final XGBoost model mentioned in the abstract is actually the XGBoost model used in model comparison. This expression has caused ambiguity. So, we delete the word "final" (see page 2, line 44, and page 2, line 44)

Changes in the text:

Page 2, line 44: "In the XGBoost model…"

Page 2, line 44: "The XGBoost model and the complete dataset …"

Comment 2: The factors that predict mortality are different depending upon the type of histology i.e. squamous cell carcinoma behaves better than adenocarcinomas. Is the proposed model valid for both types?

Reply 2: Thank you for pointing this out. The proposed model valid for both types. The feature ICD-O-3 Hist/behave has been considered, which indicates each ICD-O-3 histology code and behavior code and the respective name of that histology and behavior. The feature importance of ICD-O-3 Hist/behave shows that it has an effect on the predicting results, as displayed in Figure 3 and Figure 4 in our manuscript.

Comment 3: Most of the time dataset is divided into two halves. The first half is for training and second half is for validation. Is there any reason for not using this methodology?

Reply 3: Thanks so much for this insightful recommendation. The dataset considered in this paper does not consider the time factor. The model evaluation method used is the 5-fold cross-validation, where the total dataset is randomly divided into 5 equal sized subsamples. Among the 5 subsamples, a single subsample is held as the validation data to test the model, and the residual 4 subsamples are used as training data. The cross-validation process is repeated 5 times, with each of the 5 subsamples used exactly once as the test data. The 5 results can then be averaged to produce a single estimation, i.e., the performance measure of the model. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Changes in the text:

Page 8, line 143 to 152: "K-fold cross-validation is the most commonly resampling techniques used in evaluating ML models. The original sample is randomly divided into k equal sized subsamples. Among the k subsamples, a single subsample is held as the validation data to test the model, and the residual $k - 1$ subsamples are used as training data. The cross-validation process is repeated k times, with each of the k subsamples used exactly once as the test data. The k results can then be averaged to produce a single estimation, i.e., the performance measure of the model. In this paper, for the dataset with unequal class proportions, stratified k-fold is used, where the folds are made by preserving the percentage of samples for each class."

Comment 4: How this new model can be used in clinical practice.

Reply 4: Thank you for pointing this out. In clinical practice, we can get the patient's information according to Table 1 in our manuscript, then use the XGBoost model trained by data from SEER database to predict the 5-year survival status of the patient. The probability of 5-year survival of this patient can be computed for auxiliary diagnosis and treatment.

Comment 5: There are few mistakes and require correction.

Reply 5: Thank you very much to point out the spelling mistakes and grammatical issues in our manuscript. We have revised them point-by-point.

Changes in the text:

Page 21, line 290: “Discussion”

Page 21, line 295 to 300: “In this study, discrimination of the performance of the three newly developed variants of GBM was similar, with little variability in ROC, PRC (Figure 2) and other three metrics (Table 3). They outperform the other models, including the ANN used by Sato F (4). The predicting result (ROC_AUC=0.88) in the study of Sato F (4) is higher than that in this paper, but there are more features (199 features) used to train the model providing more information thus better predicting outcomes.”

Page 24, line 350 to 353: “Some critical factors that are strong predictors of survival and patient outcome are unavailable in the SEER database, such as methods of surgery, postoperative complications, and more importantly, radiation and chemotherapy information.”