We thank the Editor and all the Reviewers for the thorough review of our manuscript entitled "Prediction of genetic alterations from gastric cancer histopathology images using a fully automated deep learning approach". All the detailed comments from the reviewers were reflected in the revised manuscript. Below is our point-by-point response.

**Reviewer #1:**

- Although the problem and significance of this study are clearly exposed the manuscript requires significant language editing which would enable the reader to understand better the study.

The revised manuscript was re-examined and edited by an expert English editing service provider. We added 'Certificate of editing' file.

- The number of patient samples assessed are limited which as the authors suggest affected the generalizability of their developed classifiers. Although I agree that heterogeneity exists within cancer tissues from different countries, hospitals etc. and hence a mixed classifier which has trained on both the TCGS and the SSMH datasets would be more appropriate, the authors should have kept a small number of cases as their testing or validation cohort. I recommend either increasing the number of cases included in the study by specifically assessing their developed classifiers in an unseen dataset or that they divide their current dataset into a training and test/validation set.

We used different combinations of training and testing datasets.
First, the classifiers were trained for only TCGA dataset and separately tested on both TCGA and SSMH datasets. For these classifiers, 90% of TCGA data were used for training and 10% of TCGA data were used for testing. All SSMH data were used for testing as an external validation set. Therefore, it was typical setting for external validation (SSMH dataset) of deep learning model trained with another dataset (TCGA dataset).
Second, SSMH datasets were split into train and test sets and the train sets were combined with the TCGA train sets to investigate the effect of extended dataset. Therefore, new classifiers trained with both datasets could be obtained. The test sets were completely separate sets of patients compared to train sets in both TCGA and SSMH datasets.
We think the training and test/validation set splits were valid for both settings for the purpose of the study. We kept a small number of cases as their testing or validation cohort (~10%).

- Page 11 "Deep learning model": "For the tumor or mutation classifiers described below, only proper tissue patches were analyzed" what do the authors mean by proper? This sentence should be edited to make sure others can reproduce the work if they wanted to.

As the reviewer noted, our explanation was not clear. We changed the original manuscript and added some details for clarity. (Related manuscripts are as follows)

(Page 11 Line: 10-17)

A simple convolutional neural network (CNN), termed as tissue/non-tissue classifier, was trained to discriminate these various artifacts all at once. The structure of the tissue/non-tissue classifier was described in our previous study[11]. The tissue/non-tissue classifier could filter out almost 99.9% of the improper tissue patches. Then, tissue patches classified as "improper" by the tissue/non-tissue classifier were removed, and the remaining "proper" tissue patches were collected. For the tumor or mutation classifiers described below, only proper tissue patches were analyzed (Figure 1).

- Page 10 "After reviewing the quality of WSIs from the GC dataset of TCGA (TCGA-STAD), we selected slides from 25, 19, 34, 64, and 160 patients, which were confirmed to have mutations in CDH1, ERBB2, KRAS, PIK3CA, and TP53 genes, respectively. "Do these numbers represent patient numbers as well? Or are there slides which are from the same patients, just different blocks? Details regarding the selection criteria of their images should be provided and the logic behind their selections should be discussed.

The numbers (25, 19, 34, 64, and 160) indicated the number of patients because we adopted the patient-level cross validation. There were more than two slides for many patients in the TCGA datasets, maximum of 4 slides for some patients. For frozen tissue slides, one or two of these slides contained only normal tissues. We excluded normal slides and selected maximum of two tumor-containing slides per patient. The final numbers of slides were 34, 26, 50, 94, and 221 for frozen tissue slides and 27, 19, 34, 66, and 174 for FFPE tissue slides for CDH1, ERBB2, KRAS, PIK3CA, and TP53 genes, respectively. We added more details in the Methods section for clarity. (Related manuscripts are as follows)

(Page 9 Line: 16- Page 10 Line: 6)

After a carefully review of all the WSIs in the TCGA GC dataset (TCGA-STAD), we eliminated WSIs with poor scan quality and very small tumor contents. We selected slides from 25, 19, 34, 64, and 160 patients, which were confirmed to have mutations in CDH1, ERBB2, KRAS, PIK3CA, and TP53 genes, respectively. There were more than two slides for many patients in the TCGA dataset, with a maximum of four slides for some patients. However, in many cases, one or two slides contained only normal tissues. We excluded normal slides and selected a maximum of two tumor-containing slides per patient. The final number of frozen tissue slides was 34, 26, 50, 94, and 221 and that of formalin-fixed paraffin-embedded (FFPE) tissue slides was 27, 19, 34, 66, and 174 for CDH1, ERBB2, KRAS, PIK3CA, and TP53 genes, respectively.

- Page 12 "Mutation classifiers were trained separately for the selected tumor patches for frozen and FFPE tissues", details regarding the two classifiers (frozen and FFPE) should be discussed (ie. number of patients, slides, training regions etc). How many patches per image were they assessed on average? This information would be good to be added. How long were the classifiers trained for? Further details regarding their training process is crucial for reproduction.

The numbers of patients and slides were clarified as in the answer to the previous question. Because we adopted 10 fold cross-validation, the number of slides included in the training and testing were 90% and 10% of the slides for each fold. As described in the Methods section, training regions were automatically selected by the normal/tumor classifier to include only the tumor regions. We appended a new table for the number of average patches used for the training of each classifier (Supplementary Table S2). Also, average training epochs were provided in Supplementary Table S3. Furthermore, for a more clear assessment of the classifiers, we appended tables for the accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models (Table 1 and Supplementary Table S4). (Related manuscripts are as follows)

(Page 13 Line: 18-21)

The number of tissue patches used for the training of all mutation prediction models is summarized in Supplementary Table S2. The average number of training epochs for each classifier is summarized in Supplementary Table S3.

(Page 15 Line: 7-9)

In addition, the accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models with cutoff values for maximal Youden index (sensitivity + specificity - 1) were presented.

(Page 17 Line: 14-16)

For a clearer assessment of the performance of each model, the accuracy, sensitivity, specificity, and F1 score of the classification results are presented in Table 1.

(Page 19 Line: 8-10)

The accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models trained with both SSMH and TCGA datasets are presented in Supplementary Table S4.

**Reviewer #2:**

a. Author's has to highlights the major contributions of the manuscript in introduction section. Also briefly describe the flow of manuscript for the improvement in readability of the article.

As the reviewer suggested, both the major contribution and the flow of the manuscript were added to the final part of the Introduction section. (Related manuscripts are as follows)

(Page 8 Line: 20- Page 9 Line: 9)

This study investigated the feasibility of classifiers for mutations in the CDH1, ERBB2, KRAS, PIK3CA, and TP53 genes in GC tissues. First, the classifiers were trained and tested for GC tissue slides from The Cancer Genome Atlas (TCGA). The generalizability of the classifiers was tested using an external dataset. Then, new classifiers were trained for combined datasets from TCGA and external datasets to investigate the effect of the extended datasets. The results suggest that it is feasible to predict mutational status directly from tissue slides with deep learning-based classifiers. Finally, as the classifiers for KRAS, PIK3CA, and TP53 mutations for both colorectal and GC were available, we also analyzed the generalizability of the DL-based mutation classifiers trained for different cancer types.

b. Why you guys are selecting Deep learning model rather than Machine learning model. You have to also explain the reason to choose Inception model in place of other pre-trained model.

The current study belongs to our series of efforts to test the feasibility of a deep learning-based prediction system for the identification of molecular biomarkers directly from the H&E-stained tissue slides. It is well perceived that deep learning performs much better for the complex visual tasks such as the identification of molecular biomarkers directly from the H&E-stained tissue slides. We selected the Inception v3 model because it performs well on the tissue classification tests compared to other famous CNN models [20]. Furthermore, because we used the Inception v3 in the previous study to predict mutations in the colorectal cancer [11], the direct comparison could be possible. We clarified this as follows. (Related manuscripts are as follows)

(Page 12 Line: 15-18)

Thereafter, the Inception-v3 model, a widely used CNN architecture, was trained to classify the tumor patches into 'wild-type' or 'mutated' tissues, as in our previous study on mutation prediction in colorectal cancer[11].

c. In this work Authors are simply importing the pre-trained CNN model. It shows that the novelty of the work is missing in case of model development.

We did not adopt the transfer learning scheme. We fully trained the network from the beginning. We clarified that in the Methods section as follows. (Related manuscripts are as follows)

(Page 12 Line: 17-18)

We fully trained the network from the beginning and did not adopt a transfer-learning scheme.

The focus of this study was not the development of a new model but the testing of the feasibility of prediction of mutation in the gastric cancer tissue slides.

d. As we know that the deep learning model perform better for large dataset. In this case you are using very less amount of data. In this case model suffers from under fitting. So author's has requested to justify that your model is not suffering from under fitting.

Although the numbers of patients were small, the data were not. We provided a new table to clarify the numbers of tissue patches used for the training of each classifier (Supplementary Table S2). The numbers of tissue patches ranged from 168,035 for the ERBB2 gene of the TCGA FFPE Tissue Slides to 1,132,510 for the TP53 gene of the TCGA FFPE Tissue Slides. Although the numbers of tissue patches were considerable, the performance could be improved when we combined the TCGA and SSMH datasets. Therefore, the classifiers were still under fitted, and much more data should be collected. We discussed the issue. (Related manuscripts are as follows)

(Page 13 Line: 18-19)

The number of tissue patches used for the training of all mutation prediction models is summarized in Supplementary Table S2.

(Page 24 Line: 2-8)

In our opinion, the datasets are still immature for building a prominent classifier for mutation prediction. Therefore, efforts to establish a larger tissue dataset with a mutation profile will help to understand the potential of DL-based mutation prediction systems. Recently, many countries have started to build nationwide datasets of pathologic tissue

WSIs with genomic information. Therefore, we expect that the performance of DL-based mutation prediction can be greatly improved.

e. For the performance measuring, ROC is not only the sufficient approach. Specially for medical science research, you have to also perform statistical so that the significance of the designed model can be verified properly.

As the reviewer suggested the ROC curves are not sufficient for the full evaluation of the performance of the classifiers. For a more clear assessment of the classifiers, we appended tables for the accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models (Table 1 and Supplementary Table S4). (Related manuscripts are as follows)

(Page 15 Line: 7-9)

In addition, the accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models with cutoff values for maximal Youden index (sensitivity + specificity - 1) were presented.

(Page 17 Line: 14-16)

For a clearer assessment of the performance of each model, the accuracy, sensitivity, specificity, and F1 score of the classification results are presented in Table 1.

(Page 19 Line: 8-10)

The accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models trained with both SSMH and TCGA datasets are presented in Supplementary Table S4.

f. Authors have not listed the social impact of the study. You have to also mention it.

In the discussion and conclusion sections, we discussed the many impacts of the deep learning-based mutation prediction. (Related manuscripts are as follows)

(Page 20 Line: 14-21)

However, molecular tests to detect gene mutations are still not affordable for cancer patients. If cost- and time-effective alternative methods for mutation detection can be introduced, it will promote prospective clinical trials and retrospective studies to correlate the treatment response with the mutational profiles of cancer patients, which can be retrospectively obtained from clinical data and stored tissue samples. Therefore, the new cost- and time-effective methods will help to establish molecular stratification of cancer patients that can be used to determine effective treatment and improve clinical outcomes[34].

(Page 21 Line: 20- Page 22 Line: 2)

further studies for the fine molecular stratification of patients based on mutational status are ongoing[6]. DL-based mutation prediction from the tissue slides could provide valuable tools to support these efforts because the mutational status can be promptly obtained with minimal cost from the existing H&E-stained tissue slides.

We added a sentence to the final part of the conclusion section. (Related manuscripts are as follows)

(Page 25 Line: 11-12)

Furthermore, its cost- and time-effective nature could help save the medical cost and decision time for patient care.

g. How this work can be extended further?

As we discussed, the performance of deep learning-based mutation prediction still needs to be improved. We think that the first thing to do is to improve the performance of the prediction models. We recently started to participate in a Korean government-led 5-year project to build a huge dataset of cancer tissue slides with genomic data. We hope the data from the project will lead to establishing much better classifiers for the mutation prediction in cancer tissue slides including gastric cancer. We would try to improve the accuracy of the prediction models with ever-growing data from the project. We discussed the importance of larger datasets and the possibility of improved performance. (Related manuscripts are as follows)

(Page 24 Line: 2-8)

In our opinion, the datasets are still immature for building a prominent classifier for mutation prediction. Therefore, efforts to establish a larger tissue dataset with a mutation profile will help to understand the potential of DL-based mutation prediction systems. Recently, many countries have started to build nationwide datasets of pathologic tissue WSIs with genomic information. Therefore, we expect that the performance of DL-based mutation prediction can be greatly improved.

h. Authors can improve the literature work by adding some quality work like • Echle, Amelie, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. "Deep learning in cancer pathology: a new generation of clinical biomarkers." British journal of cancer 124, no. 4 (2021): 686-696. • Calderaro, Julien, and Jakob Nikolas Kather. "Artificial intelligence-based pathology for gastrointestinal and hepatobiliary cancers." Gut 70, no. 6 (2021): 1183-1193. • Coudray, Nicolas, and Aristotelis Tsirigos. "Deep learning links histology, molecular signatures and prognosis in cancer." Nature Cancer 1, no. 8 (2020): 755-757. • Bhatt, Chandradeep, Indrajeet Kumar, V. Vijayakumar, Kamred Udham Singh, and Abhishek Kumar. "The state of the art of deep learning models in medical science and their challenges." Multimedia Systems (2020): 1-15

We appended all the references to the appropriate sentences.

**Reviewer #3:**

1) Dataset composition: the dataset is highly unbalanced, so the authors randomly selected patients without mutations in order to balance the dataset. Personally, I would have selected all patients without mutations (perhaps using fewer patches from each patient) so as to train CNN on a more heterogeneous dataset.

As the reviewer noted, the tumor patches from all wild-type patients other than the test sets were randomly sampled for the training of the classifiers. The python code (https://github.com/jajman/StomachMutation/blob/main/WSI_preparation/trainvalFileSplitMutation.py) shows how the split was done. However, the current description of the methods did not clearly explain the procedures. Therefore, we revised the Methods section to describe the procedures more clearly. (Related manuscripts are as follows)

(Page 10 Line: 18- Page 11 Line: 2)

Our previous studies recognized that a DL model cannot perform optimally for both training and testing unless the dataset is forced to have similar amounts of data between classes[23]. Therefore, we limited the difference in patient numbers between the mutation and wild-type groups to less than 1.4 fold by random sampling. For example, only 35 of the 183 wild-type patients were randomly selected as the CDH1 wild-type group because there were only 25 CDH1 mutated patients. Ten-fold cross-validation was performed based on these randomly sampled wild-type patients. However, the classifiers yielded better results when the tumor patches from all wild-type patients other than the test sets were randomly sampled to match the 1.4 fold data ratio of wild-type/mutation groups for training, as this strategy could include a greater variety of tissue images. Therefore, we included all wild-type patients other than the test sets during training and randomly selected patients during testing.

2) Network training - Page 12, Line 7: ". The same label for all tumor tissue patches in a WSI as either 'wild-type' or 'mutated' were assigned based on the mutational status of the patient." Slide-level classification is very different from patch-level classification. Even if a WSI is labeled as "mutated," it

is not certain that all its tumor patches contain features related to the gene mutation. This means that the network may accept patches that are labeled as "mutated" (because they come from a "mutated" WSI) but do not actually contain any alterations. This may represent a bias during network training.

As the reviewer suggested, it is impossible to clearly discriminate the tissue image patched into 'wild-type' or 'mutated' patches considering the heterogeneity of tumor tissues. We clearly recognized the limitation and considered it as an innate limitation of this kind of task. Therefore, our goal of the current study was to understand the feasibility of mutation prediction from the H&E-stained tissue slides even though it is inherently impossible to collect perfectly labeled data. Many researchers dealt with these multiple instance learning situations and the deep learning models performed well even though the labeling was imperfect. Deep learning could learn the most discriminative features for the classification tasks from the imperfectly labeled data and successfully classified the slide into 'wild-type' or 'mutated' slides even though there were mixed patches with features of 'wild-type' or 'mutated' tissues. We think it’s the strength of deep learning to build the most apt models for the given tasks. However, more sophisticated strategies should be further developed to integrate the information from the heterogenous tissue patches to yield more precise prediction.

- Page 11, Line 3: "We divided a WSI into non-overlapping patches of 360×360 pixel tissue images at 20× magnification to detect mutational status. " How were these parameters chosen?

Because many of the WSIs in the TCGA datasets were scanned at 20× and not 40×, we decided to adopt 20× to include all possible WSIs. 360×360 pixel tissue images were determined to make the batch size for the training could be more than 100. We experimented with different tissue sizes and batch sizes during our previous studies and concluded that 360×360 pixel tissue images and the batch size of 128 are good for tissue classification tasks. Then, we used the parameters for many studies. We consider the parameters are reasonable for current GPU performance.

- Page 11,Line 13: "Morphologic features reflecting mutations in specific genes might be expressed mainly in tumor tissues rather than normal tissues." Please add at least one or two references for this sentence. - more details about the training should be provided (optimiser, transfer learning yes/no, number of epochs, early stopping criteria, etc.)

We appended two new references to the sentence (Driver and passenger mutations in cancer (PMID: 25340638, DOI: 10.1146/annurev-pathol-012414-040312), Comprehensive Characterization of Cancer Driver Genes and Mutations (PMID: 29625053, DOI: 10.1016/j.cell.2018.02.060)).

More details about the training were provided in the Methods section. (Related manuscripts are as follows)

(Page 12 Line: 17-18)

We fully trained the network from the beginning and did not adopt a transfer-learning scheme.

(Page 12 Line: 20- Page 13 Line: 4)

The Inception-v3 model was implemented using the TensorFlow DL library (http://tensorflow.org), and the network was trained with a mini-batch size of 128 and cross-entropy loss function as a loss function. For training, we used the RMSProp optimizer, with an initial learning rate of 0.1, weight decay of 0.9, momentum of 0.9, and epsilon of 1.0. Ten percent of the training slides were used as the validation dataset, and training was stopped when the loss for the validation data started to increase.

- Page 12, Line 19: "Color normalization was applied to the tissue patches to avoid the effect of stain differences.". Recent studies have shown that stain normalization is an effective preprocessing step to build reliable deep learning frameworks in digital pathology (doi: 10.1016/j.compbiomed.2020.104129, doi:10.1038/s41598-020-71420-0). At least the one reference is needed for this sentence.

We appended the two references to the sentence (The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis (PMID: 33254082, DOI: 10.1016/j.compbiomed.2020.104129), Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer (PMID: 32873856, DOI: 10.1038/s41598-020-71420-0)).

- What is the overall accuracy of the cancer detection system?

For a more clear assessment of the classifiers, we appended tables for the accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models (Table 1 and Supplementary Table S4). (Related manuscripts are as follows)

(Page 15 Line: 7-9)

In addition, the accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models with cutoff values for maximal Youden index (sensitivity + specificity - 1) were presented.

(Page 17 Line: 14-16)

For a clearer assessment of the performance of each model, the accuracy, sensitivity, specificity, and F1 score of the classification results are presented in Table 1.

(Page 19 Line: 8-10)

The accuracy, sensitivity, specificity, and F1 score of the classification results of mutation prediction models trained with both SSMH and TCGA datasets are presented in Supplementary Table S4.

- Were the same images from this study used to train the classifier? (normal/tumor classifier)

Yes, the normal/tumor classifiers were separately trained with the frozen and FFPE tissue slides of the TCGA-STAD datasets.