

Inferring Clinically Relevant Molecular Subtypes of Pancreatic Cancer from Routine Histopathology Using Deep Learning

Corresponding Author: Mr Abdul Akbar

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The Authors have done amazing work to link the RNAseq expression from the patient specimens to the Histopathology imaging results. The work is highly encouraging and has potential to make significant impact in clinical practice. Here are some of my suggestions after reviewing the manuscript.

1) It is a known fact that tumor heterogeneity affects the strength of RNAseq measurement and consequently the signature. Is the RNAseq on same slide or consecutive slide.

2) RNAseq expression values are prone to batch effects. Using two different sources PANCAN and TCGA can have an impact on the expression values. Was any batch correction performed to the data to ensure consistency of RNA expression signatures.

3) The information about selected samples used for training and validation is presented within the Supplementary tables. It will be good to have these values and the training and test accuracy presented within the main article.

4) It will be great to have a table comparing the predictions of the ground truth from RNA seq vs the model for individual datasets. Currently this is described in a section "Benchmarking against baseline reveals superior cross-cohort stability of PanSubNet "

5) An orthogonal approach such as IHC for specific Pancreatic cancer markers may also be evaluated for ground truth identification for cancer subtype

Reviewer #2

(Remarks to the Author)

Review Report

Inferring Clinically Relevant Molecular Subtypes of Pancreatic Cancer from Routine Histopathology Using Deep Learning

In this study, authors define a deep learning approach (PanSubNet) to categorize PDAC patients to classical and basal-like subtypes. Their approach reaches meaningful and possibly clinically translatable results. This is a well-written study, however there are some areas need to be addressed. Please see my comments below;

Please cite the below sentences;

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal human malignancies worldwide.

These outcomes are in stark contrast to survival rates exceeding 90% for breast and prostate cancers (citations), and approximately 65% for colorectal cancers (citation).

The therapeutic progress over the past decade has been modest.

Deep learning applied to routine hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) offers a scalable digital pathology approach to inferring molecular insights.

However, many existing methods treat molecular prediction as a hard classification problem and rely predominantly on regional correlations (citation), with limited explicit modelling of cellular morphology, spatial organization, or biological uncertainty (citation).

In Moffitt classification system, cases often fall along a spectrum spanning classical, basal-like, intermediate classical, and intermediate basal-like phenotypes.

Please cite the PRODIGE trial.

Please delete the extra space in line 71 between (and c letter.

In line 91, authors mentioned 1055 patients, thus the study comprises 987 cases. Please correct the inconsistency or explain the reason of different numbers. Also, according to that, please correct the sentence (line 285-286) of "Although this study included more than 1000 patients, additional validation in larger prospective and multi-institutional cohorts will be important to further assess clinical generalizability."

Please cite CellVit++, UNI2-h, and AttMIL in the figure legends too (at their first usage).

Please write down figure 2B with 20 intervals as in Figure 2A. It should be 0-20-40-60-80-100, instead of 60-65-70-75-80-85.

Please write down Figure 3 bolt (line 192).

Please write as metastatic subset of patients (line 199)

Please add exact p value of figure 3b into the text (line 200) and into the figure, instead of saying <0.05 only.

There are only 3 patients changed between metastatic subset of patients between RNA-seq labels and model predicted labels. Instead of saying several (line 200), please use 3 cases which is more reasonable in the sentence; Notably, 3 cases labeled as classical by.....

To examine the impact of subtype confidence on survival associations, we additionally repeated OS analyses in the full PANCAN cohort, including both high- and low-confidence cases (see Supplementary Figure S2). Across these analyses, inclusion of low-confidence cases was associated with reduced separation between survival curves compared to analyses restricted to high-confidence molecular labels. The confidence-dependent improvement in the prognostic signal suggests that internal estimates meaningfully reflect biological fidelity rather than model noise. Above argument is correct but it is not fully reflecting the whole picture. In Supplementary S2, RNA-seq labels substantially outperform the PanSubNet, which can be expected because of training with only high confidence cases. Thus, authors should emphasize this as a limitation/potential future work at the end of the study.

Please use wording of WSIs (line 235 and line 289-290), since it was already explained before.

Use K as big letter at the starting of the sentence (line 237).

The penmarks in Figure 4 and Figure 5 looks redundant in such a sophisticated deep learning architecture. Please refine these images with annotations made on digitized WSIs. Also, for these figures, please add snapshot level (such as 1x, 1.5x, 4x etc.) with emphasizing that these are H&E staining for the whole potential reader audience. In addition, it may be a good idea to add insets within these H&E's to show the highlighted attention maps areas to understand the regions are coming from biologically relevant areas (such as tumor, tumor associated stroma etc.) instead of non-relevant areas (penmarks, necrosis, artefacts etc.)

Please change the 44th citation place from the end of sentence to after "Consistent with previous reports" wording.

Please add different scanner names, if known (line 296), to highlight the variety of different platforms to support robust performance of the algorithm.

Please write all gene, mRNA, and transcript names italic. For the correction needed areas, please see below;

DNA damage repair gene expression analysis paragraph (starting at line 384)

Supplementary Methods 2 paragraph

Supplementary Methods 3 paragraph

Supplementary Methods 4 paragraph

Supplementary Methods 5 paragraph

Supplementary Methods 6 paragraph

Please use the same font within whole text for this link (it looks bigger than the others): (<https://github.com/AI4Path-Lab/PanSubNet>)

Furthermore, PanSubNet currently does not integrate clinical variables, treatment data, or immunohistochemical biomarkers, which may further improve prognostic stratification and interpretability.

Above is a very accurate argument. It can be supported with literature including citations from some use cases of multimodal AI algorithms.

In some areas, I have a suspicion of LLM-refined sentence writing, possibly for refining the grammar. Please disclose the used LLM/s within acknowledgement or in another section, if used any, at the end of the manuscript with a new paragraph.

Please add the percentages and numbers of corresponding colors in Nested donuts (Figure S1).

Supplementary Figure S2, S3 pixel/DPI quality looks less than Figure 3, Supplementary Figure 4 (which were used for similar purposes). Please use higher quality images for these.

The text is hard to read in Supplementary Figure S6 a and b. Please magnify these texts.

In supplementary methods 6, 54th and 55th citations were used (line 236, 242, and 245) which are not in the supplementary citations and are not compatible with references within the main text. Please correct this.

Citation 55 (in the main text is published in a peer-review journal/2026 (one year after arxiv publication 2025). If an article has both arxiv and peer-reviewed version, please use the peer-reviewed version of these articles. Therefore, please check whole references with that perspective.

I believe this article is the refined and completed version of citation 53. If that is the case, please mention (in the acknowledgements) that preliminary results were published in AACR 2026.

Authors used "first name et. al" style for most of the references thus citation 30 and 44 has a different format which have more than one author name. Please be consistent with citation formatting.

Reviewer #3

(Remarks to the Author)

Akbar et al presents PanSubNet, a prognostic deep learning model that differentiates classical vs basal subtype for pancreatic adenocarcinoma. Overall this is a reasonable and well performed study. I have some points of clarification that would be of benefit to the greater audience.

I am confused about the training set for PanSubNet - in the introduction; it says that it uses 1055 patients. But then in the results/methods, it says that this was exclusively trained on the high-confidence cases (171 PANCAN + 62 TCGA). That is a dramatically smaller training set. It also says later that inference on low-confidence cases were run on PanSubNet (presumably not training). If the model was trained only on the high-confidence cases, the numbers elsewhere in the manuscript should be changed as this is misleading and it seems some of the numbers quoted in PANCAN and TCGA did not even possess molecular labels.

Regarding the OS K-M analysis. I am guessing that this analysis was done by using the data from the test set folds. Was there any overlap in the data between the test sets? And did it encompass the entire PANCAN set?

The attention map figures are too low resolution to really make anything. What would be more impactful is a quantitative metric going back to the cell classifications in CellViT. For example, perhaps using the segmented lymphocytes and correlating that to classical vs basal subtype prediction. Classical PDAC has a higher degree of immune infiltration.

Why not train a multi-class classifier in which the output is: high-confidence basal vs. intermediate vs. high-confidence classical? I feel like that would be more clinically useful as the bulk of cases are intermediate it seems.

Can the authors explain the drop in performance on TCGA inference with ABMIL models, which was not seen in PanSubNet?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I am pleased by the response from the authors to the questions raised in the previous review. I think addition of cohort level information and model comparison as tables within manuscript can be highly beneficial to the reader. Additionally clarifying

the methods used to process the data can reduce the ambiguity for potential users of the model developed during this study.

Reviewer #2

(Remarks to the Author)

Thank you for successfully completing my recommendations

Reviewer #3

(Remarks to the Author)

I appreciate the authors's answers to my comments. One comment that is missing (unless I missed it) is the promised CellViT cell type classifications for the attention maps in question #3. Otherwise I have no concerns.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We extend our gratitude to the Editor and the reviewing team for their thorough evaluation and valuable recommendations. Addressing these comments has strengthened our experimental validation and refined the clarity of our manuscript. We have addressed every critique in detail below.

Reviewer #1:

1) It is a known fact that tumor heterogeneity affects the strength of RNAseq measurement and consequently the signature. Is the RNAseq on same slide or consecutive slide.

We thank the reviewer for raising this important point regarding intratumoral heterogeneity and the relationship between histology and transcriptomic measurements. We agree that this distinction is important because spatial heterogeneity within PDAC can influence bulk RNA-seq measurements and, consequently, subtype assignment. The RNA-seq data were not generated from the exact H&E slide used for image analysis; rather, histologic and molecular data were obtained from adjacent/consecutive tissue sections from the same tumor specimen, consistent with standard tissue-processing workflows. As micro-architectural and cellular features remain largely continuous across serial sections (typically 4–5 μm apart), adjacent slides preserve the representative tissue morphology and cellular composition required for robust cross-modal alignment. We have now clarified this explicitly in the dataset description and Methods so that the relationship between the histologic and transcriptomic samples is unambiguous.

2) RNAseq expression values are prone to batch effects. Using two different sources PANCAN and TCGA can have an impact on the expression values. Was any batch correction performed to the data to ensure consistency of RNA expression signatures.

We thank the reviewer for highlighting the potential influence of batch effects when integrating transcriptomic data from PANCAN and TCGA. We agree that differences in sequencing platform, library preparation, and expression distributions across cohorts must be considered carefully. To reduce scale-related variability, expression values were log-transformed and low-information features were filtered using the 75th-percentile criterion before downstream analysis. Importantly, subtype assignment was performed using single-sample gene-set enrichment analysis (ssGSEA), which is primarily rank-based within each individual sample rather than dependent on directly comparing absolute expression magnitudes across cohorts. Thus, the subtype scores are substantially less sensitive to global between-cohort scaling differences than approaches based on raw expression values. We have now expanded the transcriptomic preprocessing section to clearly describe these steps and the rationale for using ssGSEA across the two cohorts.

3) The information about selected samples used for training and validation is presented within the Supplementary tables. It will be good to have these values and the training and test accuracy presented within the main article.

We thank the reviewer for this helpful suggestion. We have now added the relevant cohort sizes and key performance measures to the main Results Section and revised Figure 2 and its accompanying text so that the internal PANCAN evaluation and independent TCGA evaluation can be interpreted directly from the main article.

4) It will be great to have a table comparing the predictions of the ground truth from RNA seq vs the model for individual datasets. Currently this is described in a section “Benchmarking against baseline reveals superior cross-cohort stability of PanSubNet ”

We agree with the reviewer. Based on the reviewer recommendation, we have added a dedicated concordance table reporting the number of evaluated cases, overall agreement, subtype-specific misclassifications, and model-confidence characteristics for discordant cases. We also revised the corresponding Results text to explicitly direct readers to this comparison rather than describing the relationship only narratively.

5) An orthogonal approach such as IHC for specific Pancreatic cancer markers may also be evaluated for ground truth identification for cancer subtype

This is an excellent suggestion. We agree that orthogonal validation using immunohistochemical markers could provide an additional and clinically familiar way to evaluate the biological fidelity of the inferred subtypes. Unfortunately, matched IHC data for appropriate subtype-associated markers are not available for the PANCAN and TCGA cases used in the present study, which prevents us from performing a sufficiently powered orthogonal IHC validation within these cohorts. We have therefore acknowledged this as an important limitation and future validation direction.

Reviewer #2:

1) Please cite the below sentences:

- ***Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal human malignancies worldwide.***
- ***These outcomes are in stark contrast to survival rates exceeding 90% for breast and prostate cancers (citations), and approximately 65% for colorectal cancers (citation).***
- ***The therapeutic progress over the past decade has been modest.***
- ***Deep learning applied to routine hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) offers a scalable digital pathology approach to inferring molecular insights.***
- ***However, many existing methods treat molecular prediction as a hard classification problem and rely predominantly on regional correlations (citation), with limited explicit modelling of cellular morphology, spatial organization, or biological uncertainty (citation).***
- ***In Moffitt classification system, cases often fall along a spectrum spanning classical, basal-like, intermediate classical, and intermediate basal-like phenotypes.***

We thank the reviewer for carefully identifying these statements that would benefit from additional supporting references. We have now added references at each of the reviewer's indicated locations and reviewed the surrounding Introduction to ensure that major factual and literature-based statements are adequately cited.

2) Please cite the *PRODIGE* trial.

We thank the reviewer for pointing this out. We agree that the *PRODIGE* trial should be cited directly when discussing the historical FOLFIRINOX outcomes used for comparison with more recent treatment trials. We have now added the appropriate *PRODIGE* citation at this location.

3) Please delete the extra space in line 71 between (and c letter.

We have removed the extra space between the parenthesis and the following character.

4) In line 91, authors mentioned 1055 patients, thus the study comprises 987 cases. Please correct the inconsistency or explain the reason of different numbers. Also, according to that, please correct the sentence (line 285-286) of “Although this study included more than 1000 patients, additional validation in larger prospective and multi-institutional cohorts will be important to further assess clinical generalizability.”

We thank the reviewer for identifying this inconsistency. We agree that our previous wording could create confusion between the total number of patients available across the source datasets, the number with paired histologic and molecular data, and the smaller high-confidence subsets used for supervised model development and validation.

To clarify, the complete source cohorts comprised 1055 patients (PANCAN, n = 846; TCGA, n = 209). However, not all patients had both evaluable H&E whole-slide images and available molecular data. After restricting the

analysis to patients with the required paired histologic and transcriptomic information, 689 patients remained, including 506 PANCAN and 183 TCGA cases. Within these molecularly characterized cohorts, 171 PANCAN and 75 TCGA cases met the predefined high-confidence subtype criteria. PanSubNet was trained and internally evaluated exclusively on the 171 high-confidence PANCAN cases, whereas the 75 high-confidence TCGA cases were reserved as an independent external validation cohort. Lower-confidence cases were used also as exploratory inference analyses and were not included in supervised model training.

We have therefore revised the Abstract, Introduction, Results, Methods, and Discussion to consistently distinguish the overall source cohorts, the 689 patients with paired H&E and molecular data, and the high-confidence subsets used for model development and validation. We have also removed the previous reference to “more than 1000 patients,” which reflected the broader available dataset rather than the patients ultimately eligible for the present analyses. The revised manuscript now reports the numbers of patients actually included at each stage of the study, which we believe provides a clearer and more accurate description of the study population.

5) Please cite CellViT++, UNI2-h, and AttMIL in the figure legends too (at their first usage).

We have now added the appropriate citations for CellViT++, UNI2-h, and AttMIL in the relevant figure legends.

6) Please in write down figure 2B with 20 intervals as in Figure 2A. It should be 0-20-40-60-80-100, instead of 60-65-70-75-80-85.

We thank the reviewer for this suggestion. We have therefore revised Figure 2B to use 20-unit intervals (0, 20, 40, 60, 80, and 100), consistent with Figure 2A.

7) Please write down Figure 3 bolt (line 192).

We thank the reviewer for catching this formatting issue. We agree and have changed “Figure 3” to bold formatting at the indicated location for consistency with the rest of the manuscript.

8) Please write as metastatic subset of patients (line 199)

We thank the reviewer for this wording suggestion. We agree that “metastatic subset of patients” is more precise in this context, and we have revised the sentence accordingly.

9) Please add exact p value of figure 3b into the text (line 200) and into the figure, instead of saying <0.05 only.

We thank the reviewer for this important suggestion. We have now replaced the threshold notation with the exact log-rank p value in both the Results text and Figure 3B.

10) There are only 3 patients changed between metastatic subset of patients between RNA-seq labels and model predicted labels. Instead of saying several (line 200), please use 3 cases which is more reasonable in the sentence; Notably, 3 cases labeled as classical by.....

We thank the reviewer for this careful observation. We have now revised the sentence to state explicitly that three cases labeled as classical by RNA-seq were predicted as basal-like by PanSubNet in the metastatic subset.

11) To examine the impact of subtype confidence on survival associations, we additionally repeated OS analyses in the full PANCAN cohort, including both high- and low-confidence cases (see Supplementary Figure S2). Across these analyses, inclusion of low-confidence cases was associated with reduced separation between survival curves compared to analyses restricted to high-confidence molecular labels. The confidence-dependent improvement in the prognostic signal suggests that internal estimates meaningfully reflect biological fidelity rather than model noise.

Above argument is correct but it is not fully reflecting the whole picture. In Supplementary S2, RNA-seq labels substantially outperform the PanSubNet, which can be expected because of training with only high confidence cases. Thus, authors should emphasize this as a limitation/potential future work at the end of the study.

We thank the reviewer for this thoughtful and important clarification. We agree that the original discussion did not fully convey the relative behavior of the RNA-seq-derived and PanSubNet-derived labels when low-confidence cases were included. In Supplementary Figure S2, RNA-seq-based subtype assignments provide stronger survival separation than PanSubNet predictions across the broader low-confidence population, which is understandable because PanSubNet was deliberately trained only on molecularly well-defined, high-confidence cases. We have now explicitly acknowledged this limitation in the Discussion and emphasized that future work should include larger cohorts with intermediate phenotypes and modeling strategies specifically designed to learn the full subtype continuum.

12) Please use wording of WSIs (line 235 and line 289-290), since it was already explained before.

We thank the reviewer for noting this terminology issue. We have now replaced the repeated phrase with “WSIs” at the indicated locations.

13) Use K as big letter at the starting of the sentence (line 237).

We have corrected the sentence so that it begins with a capital “K.”

14) The penmarks in Figure 4 and Figure 5 looks redundant in such a sophisticated deep learning architecture. Please refine these images with annotations made on digitized WSIs. Also, for these figures, please add snapshot level (such as 1x, 1.5x, 4x etc.) with emphasizing that these are H&E staining for the whole potential reader audience. In addition, it may be a good idea to add insets within these H&E's to show the highlighted attention maps areas to understand the regions are coming from biologically relevant areas (such as tumor, tumor associated stroma etc.) instead of non-relevant areas (penmarks, necrosis, artefacts etc.)

We thank the reviewer for this very helpful suggestion. We agree that the interpretability figures should allow readers to evaluate whether highly attended regions correspond to biologically meaningful tissue rather than artifacts or slide markings. We have therefore refined Figures 4 and 5 using annotations directly on the digitized H&E WSIs, removed visually distracting pen-mark emphasis, added the corresponding viewing magnification/snapshot level, and incorporated higher-magnification insets of representative high-attention regions. These revisions make it easier to appreciate the histologic context of the highlighted regions and to distinguish tumor- and microenvironment-associated morphology from necrosis, artifacts, or other non-informative areas.

15) Please change the 44th citation place from the end of sentence to after “Consistent with previous reports” wording.

We thank the reviewer for pointing this out. We agree that the citation placement should make clear which statement it supports. We have moved citation 44 to immediately follow the phrase “Consistent with previous reports,” as suggested.

16) Please add different scanner names, if known (line 296), to highlight the variety of different platforms to support robust performance of the algorithm.

We thank the reviewer for this helpful suggestion. We agree that scanner heterogeneity is relevant when assessing cross-cohort robustness. We reviewed the available metadata; however, scanner manufacturer/model information was not consistently available for all PANCAN and TCGA cases. We have therefore clarified in the

Methods that the cohorts comprise multi-institutional WSIs acquired using heterogeneous scanning platforms, while avoiding assignment of specific scanner models where metadata could not be reliably verified.

17) Please write all gene, mRNA, and transcript names italic. For the correction needed areas, please see below;

- ***DNA damage repair gene expression analysis paragraph (starting at line 384)***
- ***Supplementary Methods 2 paragraph***
- ***Supplementary Methods 3 paragraph***
- ***Supplementary Methods 4 paragraph***
- ***Supplementary Methods 5 paragraph***
- ***Supplementary Methods 6 paragraph***

We thank the reviewer for this careful formatting review. We agree that gene, mRNA, and transcript symbols should follow consistent scientific nomenclature throughout the manuscript and Supplementary Material. We have therefore reviewed the indicated sections, including the DDR analysis and Supplementary Methods 2–6, and italicized the relevant gene and transcript symbols throughout.

18) Please use the same font within whole text for this link (it looks bigger than the others): (<https://github.com/AI4Path-Lab/PanSubNet>)

We thank the reviewer for noting this formatting inconsistency. We agree and have reformatted the GitHub link so that its font and size are consistent with the surrounding manuscript text.

19) Furthermore, PanSubNet currently does not integrate clinical variables, treatment data, or immunohistochemical biomarkers, which may further improve prognostic stratification and interpretability.

Above is a very accurate argument. It can be supported with literature including citations from some use cases of multimodal AI algorithms.

We agree that the current absence of clinical variables, treatment information, and immunohistochemical biomarkers represents an important limitation of the present image-only framework and that multimodal integration could improve both prognostic stratification and biological interpretability. We have therefore expanded this limitation in the Discussion and added relevant literature illustrating the value of multimodal AI approaches that integrate histopathology with clinical and molecular information. We also highlight multimodal extension of PanSubNet as an important direction for future work.

20) In some areas, I have a suspicion of LLM-refined sentence writing, possibly for refining the grammar. Please disclose the used LLM/s within acknowledgement or in another section, if used any, at the end of the manuscript with a new paragraph.

We thank the reviewer for raising this point and agree that the use of generative AI tools in manuscript preparation should be transparently disclosed. ChatGPT was used only to assist with language editing, grammar, and refinement of written text; it was not used to generate, analyze, or interpret the study data or to perform any statistical analyses. We have now added a disclosure statement at the end of the manuscript describing this use.

21) Please add the percentages and numbers of corresponding colors in Nested donuts (Figure S1).

We agree that the nested donut plots are easier to interpret when the absolute sample counts and corresponding percentages are shown directly. We have therefore revised Supplementary Figure S1 to display both the number and percentage represented by each color/category.

22) Supplementary Figure S2, S3 pixel/DPI quality looks less than Figure 3, Supplementary Figure 4 (which were used for similar purposes). Please use higher quality images for these.

We agree that the resolution should be consistent across all survival and molecular characterization figures. We have therefore regenerated Supplementary Figures S2 and S3 at higher resolution to improve visual quality and readability.

23) The text is hard to read in Supplementary Figure S6 a and b. Please magnify these texts.

We agree that the text in Supplementary Figure S6A and S6B was difficult to read at the previous scale. We have enlarged the labels and regenerated these panels at higher resolution to improve readability.

24) In supplementary methods 6, 54th and 55th citations were used (line 236, 242, and 245) which are not in the supplementary citations and are not compatible with references within the main text. Please correct this.

We thank the reviewer for carefully identifying this reference-numbering inconsistency. We agree that the Supplementary Methods references should be internally consistent and compatible with the reference list. We have corrected the citations in Supplementary Methods 6 and reviewed the surrounding supplementary references to ensure that the numbering is accurate.

25) Citation 55 (in the main text is published in a peer-review journal/2026 (one year after arxiv publication 2025). If an article has both arxiv and peer-reviewed version, please use the peer-reviewed version of these articles. Therefore, please check whole references with that perspective.

We thank the reviewer for bringing this to our attention. We agree that, when a peer-reviewed version of a previously cited preprint is available, the peer-reviewed publication should be preferentially cited. We have replaced the indicated arXiv citation with its peer-reviewed 2026 publication and reviewed the full reference list for similar instances.

26) I believe this article is the refined and completed version of citation 53. If that is the case, please mention (in the acknowledgements) that preliminary results were published in AACR 2026.

We thank the reviewer for recognizing the relationship between the preliminary and current work. We agree that this should be disclosed transparently. We have therefore added a statement in the Acknowledgements noting that preliminary results related to this study were presented/published in abstract form at AACR 2026.

27) Authors used “first name et. al” style for most of the references thus citation 30 and 44 has a different format which have more than one author name. Please be consistent with citation formatting.

We thank the reviewer for identifying this inconsistency. We agree that the reference list should follow a uniform citation style. We have therefore standardized the formatting of the indicated references, including citations 30 and 44, and reviewed the full reference list for consistency.

Reviewer #3:

1) I am confused about the training set for PabSubNet - in the introduction; it says that it uses 1055 patients. But then in the results/methods, it says that this was exclusively trained on the high-confidence cases (171 PANCAN + 62 TCGA). That is a dramatically smaller training set. It also says later that inference on low-confidence cases were run on PanSubNet (presumably not training). If the model was trained only on the high-confidence cases, the numbers elsewhere in the manuscript should be changed as this is misleading and it seems some of the numbers quoted in PANCAN and TCGA did not even possess molecular labels.

We thank the reviewer for identifying this inconsistency. We agree that our previous wording could create confusion between the total number of patients available across the source datasets, the number with paired histologic and molecular data, and the smaller high-confidence subsets used for supervised model development and validation.

To clarify, the complete source cohorts comprised 1055 patients (PANCAN, n = 846; TCGA, n = 209). However, not all patients had both evaluable H&E whole-slide images and available molecular data. After restricting the analysis to patients with the required paired histologic and transcriptomic information, 689 patients remained, including 506 PANCAN and 183 TCGA cases. Within these molecularly characterized cohorts, 171 PANCAN and 75 TCGA cases met the predefined high-confidence subtype criteria. PanSubNet was trained and internally evaluated exclusively on the 171 high-confidence PANCAN cases, whereas the 75 high-confidence TCGA cases were reserved as an independent external validation cohort. Lower-confidence cases were used also as exploratory inference analyses and were not included in supervised model training.

We have therefore revised the Abstract, Introduction, Results, Methods, and Discussion to consistently distinguish the overall source cohorts, the 689 patients with paired H&E and molecular data, and the high-confidence subsets used for model development and validation. We have also removed the previous reference to “more than 1000 patients,” which reflected the broader available dataset rather than the patients ultimately eligible for the present analyses. The revised manuscript now reports the numbers of patients actually included at each stage of the study, which we believe provides a clearer and more accurate description of the study population.

2) Regarding the OS K-M analysis. I am guessing that this analysis was done by using the data from the test set folds. Was there any overlap in the data between the test sets? And did it encompass the entire PANCAN set?

We thank the reviewer for this very important question. We agree that the construction of the Kaplan–Meier analysis must be clear to ensure that survival associations are not influenced by training-set predictions or overlap between test sets. For the five-fold cross-validation analysis, each PANCAN case appeared in the held-out test set exactly once. We therefore concatenated only the out-of-fold predictions from the five test folds to obtain one independent prediction per patient for the survival analysis. No patient contributed predictions from a fold in which that patient had been used for model training, and the test folds were mutually exclusive. Consequently, the pooled out-of-fold predictions collectively cover the full evaluated PANCAN subset without duplication while preserving strict separation between training and test samples. We have now clarified this procedure in the Methods and survival-analysis description.

3) The attention map figures are too low resolution to really make anything. What would be more impactful is a quantitative metric going back to the cell classifications in CellViT. For example, perhaps using the segmented lymphocytes and correlating that to classical vs basal subtype prediction. Classical PDAC has a higher degree of immune infiltration.

We thank the reviewer for this excellent suggestion. We agree that quantitative cellular analysis provides a stronger biological complement to qualitative attention maps and can directly test whether the model-associated regions reflect known differences in the PDAC microenvironment. We have therefore revised the attention maps using the CellViT++ cell classifications to quantify subtype-associated cellular compositions. We have also revised the interpretability section to present these quantitative cellular findings alongside the refined attention-map visualizations, providing a more direct connection between PanSubNet predictions and biologically meaningful cellular morphology.

4) Why not train a multi-class classifier in which the output is: high-confidence basal vs. intermediate vs. high-confidence classical? I feel like that would be more clinically useful as the bulk of cases are intermediate it seems.

We thank the reviewer for this thoughtful question, and we agree that explicitly identifying intermediate tumors is clinically attractive because a substantial fraction of PDACs occupy the transcriptional space between the classical and basal-like extremes. Our decision to train the primary model as a binary classifier on high-confidence cases was intended to minimize label noise rather than to imply that intermediate biology is unimportant. The Moffitt subtype axis is continuous, and the boundaries used to define an “intermediate” class are not a naturally discrete biological category; cases within this region can vary substantially and may lie much closer to either the classical or basal-like state. Treating all such cases as a single third class would therefore force the model to learn an internally heterogeneous category whose definition is sensitive to dataset-specific thresholds, potentially reducing both biological fidelity and cross-cohort generalizability. We instead trained PanSubNet on the molecularly clean extremes and then evaluated whether the learned representation generalizes across the full subtype spectrum. Importantly, the model retained measurable discriminatory ability when applied to the broader PANCAN cohort, and an auxiliary analysis showed that the learned representations also contain information related to subtype confidence. We have expanded the manuscript to clarify this rationale and now explicitly identify direct modeling of the continuous/intermediate phenotype as an important future direction, particularly as larger molecularly annotated cohorts become available.

5) Can the authors explain the drop in performance on TCGA inference with ABMIL models, which was not seen in PanSubNet?

We agree that the stronger internal performance of AttMIL and its larger degradation on TCGA warrants careful interpretation rather than being attributed to a single definitive cause. One plausible explanation is that a conventional attention-based MIL model can achieve high internal discrimination by concentrating on a relatively small set of cohort-specific patch-level correlates, some of which may be sensitive to differences in staining, tissue preparation, scanner characteristics, or case composition across institutions. PanSubNet, in contrast, was designed to integrate complementary cellular and tissue-scale representations before slide-level aggregation, encouraging the prediction to depend on a broader set of morphologic signals rather than a single patch representation. This may contribute to the more balanced sensitivity and specificity and the smaller cross-cohort performance shift observed on TCGA. Nevertheless, we agree that the present data cannot establish the precise mechanistic reason for AttMIL's larger external degradation. We have therefore revised the Discussion to present this as a plausible interpretation rather than a definitive conclusion and to emphasize that additional multi-institutional validation and controlled domain-shift analyses will be needed to determine which architectural components are most responsible for the observed generalizability.