medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Deep learning for subtypes identification of pure seminoma of the testis

Kirill E. Medvedev[1,*], Paul H. Acosta[2], Liwei Jia[3], Nick V. Grishin[1,4]

[1]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[2]Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[3]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[4]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

***** Corresponding author

E-mail: Kirill.Medvedev@UTSouthwestern.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Abstract

The most critical step in the clinical diagnosis workflow is the pathological evaluation of each tumor sample. Deep learning is a powerful approach that is widely used to enhance diagnostic accuracy and streamline the diagnosis process. In our previous study using omics data, we identified two distinct subtypes of pure seminoma. Seminoma is the most common histological type of testicular germ cell tumors (TGCTs). Here we developed a deep learning decision making tool for the identification of seminoma subtypes using histopathological slides. We used all available slides for pure seminoma samples from The Cancer Genome Atlas (TCGA). The developed model showed an area under the ROC curve of 0.896. Our model not only confirms the presence of two distinct subtypes within pure seminoma but also unveils the presence of morphological differences between them that are imperceptible to the human eye.

Keywords: bioinformatics, computational biology, deep learning, seminoma, subtypes

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Introduction

Testicular seminoma is the most prevalent histological subtype of testicular germ cell tumors (TGCT), accounting for the highest incidence rate among all types of testicular cancer [1]. TGCTs are the most frequent type of solid cancer affecting men between the ages of 15 and 44 [1] and rank second among adult cancers in terms of life years lost per person dying of cancer [2]. The treatment protocol for seminoma typically includes orchiectomy followed by either platinum-based chemotherapy utilizing cisplatin or radiation therapy [3]. While current treatments for seminoma have high efficacy and survival rates for patients, they also carry the risk of around 40 severe and potentially life-threatening long-term side effects, such as infertility, neurotoxicity, hypercholesterolaemia, secondary cancers and death [4]. The presence of elevated platinum concentrations from chemotherapy in the bloodstream can persist at levels up to 1,000 times higher than the norm for a duration of 20 years, potentially contributing to various long-term effects [5]. Prolonged exposure to elevated platinum levels can result in vascular damage and is highly likely to be linked with the onset of neuropathy [6] and cardiovascular diseases [7]. After undergoing chemotherapy, patients with TGCT exhibited a decrease of 3.6 dB in hearing for each additional 100 mg/m2 of cumulative cisplatin dose [8,9]. Relapse occurs in approximately 20% of seminoma cases, and the underlying reasons for this phenomenon remain unclear [10], however there are several well-known seminoma risk factors such as rete testis, lymphovascular invasion, cryptorchidism, mutations in KRAS and KIT genes [11,12]. Patients experiencing a relapse will receive further treatment involving chemotherapy and radiation therapy, which intensify the side effects considerably. Recently we discovered two distinct subtypes of pure seminoma of the testis based on omics data [13,14]. Two identified seminoma subtypes revealed significant differences in the rates of loss of heterozygosity, the level of expression of lncRNA associated with cisplatin resistance,

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

the activity of double stranded DNA breaks repair mechanisms and the pluripotency stage. Seminoma subtype 1 exhibits a higher pluripotency state, while subtype 2 reveals attributes of reprogramming into non-seminomatous lineages of TGCT, which are more aggressive and require higher dose of chemotherapy drugs [3]. We showed that subtype 1 of seminoma, which is less differentiated, exhibits an immune microenvironment characterized by a significantly lower immune score and a larger fraction of neutrophils [14]. These features are indicative of the immune microenvironment at an early developmental stage. Moreover, subtype 2 revealed the overexpression of genes related to the senescence-associated secretory phenotype, which might be one of the reasons for seminoma immunotherapy failure [14]. Therefore, we suggested that seminoma subtype 2 might require an adjustment to its treatment strategy. The development of subtype-specific therapy for seminoma can reduce the risk of chemotherapy overtreatment in TGCT patients and enhance the quality of life for TGCT survivors.

Deep learning (DL) is a powerful tool capable of extracting previously hidden information directly from routine histopathology images of cancer tissue, simplifying, speeding up, and automating clinical decision-making [15]. The performance of modern DL methods applied to pathological data often exceeds that of human pathologists [15]. Most pathologists work under conditions of an extreme work overload [16]. An overworked pathologist can result in the misinterpretation of pathological data that affects patients' health and quality of life. DL applications aim to simplify and speed up routine pathological workflows and reduce pathologists' overload burden.

Here, we have developed a DL-based approach to examine potential histopathologic differences between seminoma subtypes that were previously identified using omics data. Additionally, our goal is to utilize this approach to detect and classify these subtypes based on histopathological slides. Our findings demonstrate that pure seminoma subtypes cannot be classified solely based on

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

77 histopathological features. However, the developed DL-based model revealed histopathological
78 differences between these subtypes, as indicated by the area under the ROC curve (AUC) values.
79

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Materials and Methods

Data set preparation

We used all hematoxylin and eosin (H&E) histopathological slides available at The Cancer Genome Atlas (TCGA) data portal (TCGA-TGCT study) for 64 pure seminoma, which comprised 156 whole slide images (WSIs). Based on our previous study, we assigned 40 out of 64 samples to seminoma subtype 1 (101 WSIs) and the remaining 24 samples to seminoma subtype 2 (55 WSIs) [13]. Pure seminoma regions of interest (ROIs) were designatedand verified for each WSI by a genitourinary specialized pathologist using Aperio ImageScope version 12.1. During our analysis of pure seminoma H&E slides from the TCGA portal, we identified two samples (TCGA-2G-AAG9, TCGA-2G-AAH0), initially reported as pure seminoma. However, upon further examination, they should be reclassified as mixed GCT since they contain other types of GCT (teratoma and embryonal carcinoma) in addition to seminoma (Fig.1). Consequently, these cases were removed from our data set.

Verified ROIs were subsequently split into smaller tiles (300x300 pixels) at a 20X magnification with a 50% overlap using DeepPath package [17] (Fig. 2). Tiles that contained more than 20% of background were removed. We conducted a manual check and excluded tiles of poor quality that contained out of focus images and defects, such as scratches, dirt and folded tissue. TGCTs are relatively uncommon compared to other cancers [2], and therefore histopathological data availability is limited. Thus, image data augmentation technique was applied to the dataset of tiles to create synthetic variations of the images and expand the training dataset. We used the following augmentations from the Imgaug library [18]: random rotation by 0°, 90°, 180°, 270°, random increase and decrease of contrast, brightness and saturation.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Training the model

We conducted all computational experiments at the BioHPC computing facility (Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, TX, USA). TensorFlow package [19] version 2.5.0 was used for developing DL model based on Convolution Neural Networks (CNNs) method. The software stack for GPU acceleration included CUDA 11.2 and cuDNN 8.1. We employed a convolutional neural network with MobileNet [20] architecture that includes 85 convolutional layers. During the training top 29 convolutional layers were kept frozen (fixed). We used sigmoid activation and the Binary Cross Entropy loss function, which is used when there are only two label classes (seminoma subtypes 1 and 2). The Adam optimizer was selected due to its superior performance in terms of both speed of convergence and accuracy [21] with a learning rate of 0.001. Due to imbalanced dataset we used subtype (class) weights calculated as follows: $w_1 = (1 / s_1) * (total / 2.0)$ and $w_2 = (1 / s_2) * (total / 2.0)$, where s_1, s_2 – number of tiles for subtypes 1 and 2, and total is overall number of tiles. The neural network was initialized from ImageNet-pretrained [22] weights. Model training was performed for 20 epochs with 3-fold cross-validation. Tiles belonging to a particular sample were included only in one subset of data – either training or validation.

Statistical analysis

We used the area under receiver operating characteristics (ROC, AUC) curve and accuracy as evaluation metrics to measure the tile-level and sample-level performance of the developed model. The ROC curve was defined as false-positive rate (1-specificity) on the x-axis versus true positive rate (sensitivity or recall) on the y-axis. Specificity = TN / (TN+FP), Sensitivity = TP / (TP+FN), Accuracy = (TP+TN) / (TP+FP+TN+FN), where FP, FN, TP and TN are false positives, false negatives, true positives and true negatives, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Nuclei segmentation

Nuclei segmentation of seminoma tiles was conducted using TIA Toolbox 1.4.1 [23]. We applied the HoVer-Net model [24] that has been already trained on the PanNuke dataset [25] and incorporated in the TIA Toolbox. The calculation of nuclei size was performed using the Python library scikit-image [26].

Results and Discussion

Overview of the whole experiment is shown on Figure 3. The performance of the model was evaluated using the area under the ROC curve metric (Fig. 4B) and confusion matrices (Fig. 4C). The developed model showed highest AUC = 0.896 (Fig. 4B). Trained model for identification of pure seminoma subtypes is available in open access at GitHub (https://github.com/kirmedvedev/seminoma-subtypes).

Every sample of solid tumor undergoes a detailed analysis by a professional pathologist, which includes verifying the presence of cancer tissue and annotating tumor regions. This is a crucial process in the clinical diagnosis routine. Inaccuracies in pathology reports can critically affect the quality of patient care. An audit of pathology reports showed that if a sample size of 50 gives a sample error rate of 2% there is a 95% probability that the true error rate is up to 10.9% [27]. Moreover, up to one-third of clinicians do not always understand pathology reports, leading tomisinterpretation and uncertainty in clinical diagnosis [28]. DL approaches applied to histopathological slide images aim to speed up the diagnosis significantly and simplify their implementation into the clinical workflow. TGCTs and seminoma histopathology images, in particular, have not been extensively studied using DL method, and very limited studies are available nowadays. DL approached were previously applied to TGCTs WSIs for detecting tumor-

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

149 infiltrating lymphocytes [29], detecting lymphovascular invasion [30] and developing tumor/normal
150 classifier [31].

151 In this report, we present our first version of the DL decision making tool for the identification of
152 pure seminoma subtypes using histopathological slides. We hypothesize that considering
153 seminoma subtypes during the development of a treatment strategy may improve its clinical
154 management, and the implementation of the developed model will enhance diagnostic accuracy
155 and reduce potential errors. This is especially crucial when subtypes cannot be distinguished by a
156 pathologist which is the case with pure seminoma. The developed model showed the capability to
157 distinguish pure seminoma subtypes (Table 1),confirming our previous findings [13,14]. This also
158 indicates the presence of morphological differences between seminoma subtypes. We believe that
159 the morphological differences may be due to the difference in the immune microenvironment
160 between the two subtypes. Previously, using deconvolution methods for bulk RNA-seq data of the
161 seminoma subtypes from TCGA, we showed that the neutrophil fraction is significantly higher for
162 subtype 1 [14]. Moreover, according to TCGA clinical data, seminoma subtype 2 revealed an
163 increased occurrence of lymphovascular invasion, with a rate of 43% compared to 25% for subtype
164 1. We also conducted nuclei segmentation and calculated nuclei sizes for both subtypes. Our
165 analysis revealed no significant differences in nuclei size distributions between the seminoma
166 subtypes (Fig. 5A, B).

167 However, the accuracy of identifying subtype 1 samples by the developed model is higher than of
168 subtype 2 samples. This could be due to certain limitations of the model. First, the current model
169 was developed using a limited dataset. Second, only one architecture type of CNN was tested. In
170 future work, this model should be verified using an expanded dataset and several addition CNN
171 architecture types.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Table 1. Prediction of seminoma subtypes using developed DL model. S1 – subtype 1, S2 – subtype 2.

Sample ID	Transcriptomic subtype	DL model prediction	Sample ID	Transcriptomic subtype	DL model prediction
TCGA-XY-A9T9	S1	S1	TCGA-2G-AAH3	S1	S1
TCGA-WZ-A7V4	S1	S1	TCGA-SB-A6J6	S1	S1
TCGA-2G-AAEX	S1	S1	TCGA-2G-AAHP	S1	S1
TCGA-2G-AAF6	S1	S1	TCGA-VF-A8AB	S2	S1
TCGA-S6-A8JX	S1	S1	TCGA-ZM-AA05	S2	S2
TCGA-XY-A89B	S1	S1	TCGA-XE-AANR	S2	S2
TCGA-2G-AAFG	S1	S2	TCGA-ZM-AA0D	S2	S2
TCGA-2G-AAH8	S1	S1	TCGA-2G-AAHN	S2	S2
TCGA-2G-AAF1	S1	S1	TCGA-4K-AAAL	S2	S1
TCGA-WZ-A7V3	S1	S1	TCGA-VF-A8AE	S2	S1
TCGA-S6-A8JY	S1	S1	TCGA-ZM-AA0B	S2	S2
TCGA-VF-A8AA	S1	S1	TCGA-XE-AAOF	S2	S1
TCGA-WZ-A7V5	S1	S1	TCGA-ZM-AA0F	S2	S2
TCGA-SO-A8JP	S1	S1	TCGA-2G-AAHA	S2	S1
TCGA-XE-A8H4	S1	S1	TCGA-2G-AAHT	S2	S1
TCGA-XE-A8H5	S1	S1	TCGA-4K-AA1H	S2	S2
TCGA-XE-A9SE	S1	S2	TCGA-VF-A8A9	S2	S2
TCGA-XE-AANJ	S1	S1	TCGA-VF-A8AC	S2	S1
TCGA-XE-AANV	S1	S1	TCGA-XE-AAO6	S2	S2
TCGA-XE-AAO3	S1	S1	TCGA-XE-AAOL	S2	S2
TCGA-YU-A90Q	S1	S2	TCGA-YU-A90S	S2	S1
TCGA-YU-A90W	S1	S1	TCGA-ZM-AA06	S2	S2
TCGA-YU-A912	S1	S2	TCGA-ZM-AA0E	S2	S2
TCGA-S6-A8JW	S1	S1	TCGA-ZM-AA0H	S2	S1
TCGA-2G-AAEW	S1	S1	TCGA-ZM-AA0N	S2	S2
TCGA-2G-AAF4	S1	S1	TCGA-2G-AAHL	S2	S1
TCGA-4K-AA1I	S1	S2	TCGA-4K-AA1G	S2	S2
TCGA-2G-AAFE	S1	S1			

Conclusion

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

In this study we developed a DL-based model to investigate the presence of histopathological distinctions between two previously identified subtypes of pure seminoma, which were initially characterized using omics data. The objective was to provide further evidence supporting the existence of seminoma subtypes. The results of our analysis revealed histopathological differences between the two subtypes of pure seminoma. These findings provide additional confirmation and support the notion that seminoma can be further stratified into distinct subtypes. These results highlight the potential of histopathological analysis as a complementary tool in subtype classification, offering additional insights alongside other omics-based approaches. However, our study also provides evidence suggesting that pure seminoma subtypes cannot be reliably classified based solely on histopathological features. Despite the observed histopathological differences between the subtypes, these distinctions alone are not sufficient for accurate subtype classification.

Funding

The study is supported by the grants from the National Institute of General Medical Sciences of the National Institutes of Health GM127390 (to N.V.G.), the Welch Foundation I-1505 (to N.V.G.), and the National Science Foundation DBI 2224128 (to N.V.G.).

Acknowledgements

Authors are grateful to Dr. Satwik Rajaram for helpful discussions. Authors are grateful to TCGA data portal for providing access to TGCT datasets. This research was supported in part by the computational resources provided by the BioHPC computing facility located in the Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, TX. URL: https://portal.biohpc.swmed.edu

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Conflict of interest statement

The authors declare that there are no competing interests associated with the manuscript.

Ethics statement

Not applicable.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figures legends

Figure 1. Slide images of samples containing additional types of GCT. (A) TCGA-2G-AAG9. **(B)** TCGA-2G-AAH0. Teratoma tissue is shown in green, embryonal carcinoma in cyan.

Figure 2. Extraction of tiles from annotated whole slide images (WSI) of two subtypes of pure seminoma.

Figure 3. Overview of dataset preparation, training and validation process.

Figure 4. Validation statistics. (A) Validation accuracy. **(B)** Receiver operating characteristic (ROC) curves for validation set. **(C)** Normalized confusion matrices.

Figure 5. Nuclei segmentation results. (A) Visualization of nuclei segmentation results. **(B)** Comparison of seminoma cell nuclei size distribution between two subtypes.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

References

1. Ghazarian AA, Trabert B, Devesa SS, McGlynn KA. Recent trends in the incidence of testicular germ cell tumors in the United States. *Andrology*. Jan 2015;3(1):13-8. doi:10.1111/andr.288

2. Data from: Surveillance, Epidemiology, and End Results (SEER) Program SEER*Stat Database: Incidence - SEER Research Data, 8 Registries, Nov 2021 Sub (1975-2020) - Linked To County Attributes - Time Dependent (1990-2020) Income/Rurality, 1969-2020 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2023, based on the November 2022 submission.

3. Shin YS, Kim HJ. Current management of testicular cancer. *Korean J Urol*. Jan 2013;54(1):2-10. doi:10.4111/kju.2013.54.1.2

4. Chovanec M, Lauritsen J, Bandak M, et al. Late adverse effects and quality of life in survivors of testicular germ cell tumour. *Nat Rev Urol*. Apr 2021;18(4):227-245. doi:10.1038/s41585-021-00440-w

5. Gerl A, Schierl R. Urinary excretion of platinum in chemotherapy-treated long-term survivors of testicular cancer. *Acta Oncol*. 2000;39(4):519-22. doi:10.1080/028418600750013447

6. Sprauten M, Darrah TH, Peterson DR, et al. Impact of long-term serum platinum concentrations on neuro- and ototoxicity in Cisplatin-treated survivors of testicular cancer. *J Clin Oncol*. Jan 20 2012;30(3):300-7. doi:10.1200/JCO.2011.37.4025

7. Boer H, Proost JH, Nuver J, et al. Long-term exposure to circulating platinum is associated with late effects of treatment in testicular cancer survivors. *Ann Oncol*. Nov 2015;26(11):2305-10. doi:10.1093/annonc/mdv369

8. Haugnes HS, Stenklev NC, Brydoy M, et al. Hearing loss before and after cisplatin-based chemotherapy in testicular cancer survivors: a longitudinal study. *Acta Oncol*. Aug 2018;57(8):1075-1083. doi:10.1080/0284186X.2018.1433323

9. Frisina RD, Wheeler HE, Fossa SD, et al. Comprehensive Audiometric Analysis of Hearing Impairment and Tinnitus After Cisplatin-Based Chemotherapy in Survivors of Adult-Onset Cancer. *J Clin Oncol*. Aug 10 2016;34(23):2712-20. doi:10.1200/JCO.2016.66.8822

10. Warde P, Specht L, Horwich A, et al. Prognostic factors for relapse in stage I seminoma managed by surveillance: a pooled analysis. *J Clin Oncol*. Nov 15 2002;20(22):4448-52. doi:10.1200/JCO.2002.01.038

11. Boormans JL, Mayor de Castro J, Marconi L, et al. Testicular Tumour Size and Rete Testis Invasion as Prognostic Factors for the Risk of Relapse of Clinical Stage I Seminoma Testis Patients Under Surveillance: a Systematic Review by the Testicular Cancer Guidelines Panel. *Eur Urol*. Mar 2018;73(3):394-405. doi:10.1016/j.eururo.2017.09.025

12. Fukawa T, Kanayama HO. Current knowledge of risk factors for testicular germ cell tumors. *Int J Urol*. Apr 2018;25(4):337-344. doi:10.1111/iju.13519

13. Medvedev KE, Savelyeva AV, Chen KS, Bagrodia A, Jia L, Grishin NV. Integrated Molecular Analysis Reveals 2 Distinct Subtypes of Pure Seminoma of the Testis. *Cancer Inform*. 2022;21:11769351221132634. doi:10.1177/11769351221132634

14. Savelyeva AV, Medvedev KE. Seminoma subtypes differ in the organization and functional state of the immune microenvironment. *3 Biotech*. Mar 2023;13(3):110. doi:10.1007/s13205-023-03530-1

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

15. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer*. Feb 2021;124(4):686-696. doi:10.1038/s41416-020-01122-x
16. Fritzsche FR, Ramach C, Soldini D, et al. Occupational health risks of pathologists--results from a nationwide online questionnaire in Switzerland. *BMC Public Health*. Dec 6 2012;12:1054. doi:10.1186/1471-2458-12-1054
17. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. Oct 2018;24(10):1559-1567. doi:10.1038/s41591-018-0177-5
18. *imgaug*. 2020. Accessed accessed 01-Feb-2022. https://github.com/aleju/imgaug
19. Abadi M BP, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. TensorFlow: a system for Large-Scale machine learning. presented at: In12th USENIX symposium on operating systems design and implementation (OSDI 16); 2016;
20. Andrew G. Howard MZ, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv*. 2017;doi:https://doi.org/10.48550/arXiv.1704.04861
21. Kingma DP BJ. Adam: A method for stochastic optimization. *arXiv*. 2014;doi:https://doi.org/10.48550/arXiv.1412.6980
22. Deng J DW, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. presented at: In 2009 IEEE conference on computer vision and pattern recognition 2009 Jun 20; 2009;
23. Pocock J, Graham S, Vu QD, et al. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Commun Med (Lond)*. 2022;2:120. doi:10.1038/s43856-022-00186-5
24. Graham S, Vu QD, Raza SEA, et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal*. Dec 2019;58:101563. doi:10.1016/j.media.2019.101563
25. Jevgenij Gamper NAK, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, Nasir Rajpoot. PanNuke Dataset Extension, Insights and Baselines. *arXiv*. 2020;doi:10.48550/arXiv.2003.10778
26. van der Walt S, Schonberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ*. 2014;2:e453. doi:10.7717/peerj.453
27. Ramsay AD. Errors in histopathology reporting: detection and avoidance. *Histopathology*. Jun 1999;34(6):481-90. doi:10.1046/j.1365-2559.1999.00719.x
28. Mirham L HJ, Yousef GM. Addressing the diagnostic miscommunication in pathology: old challenges and innovative solutions. *American Journal of Clinical Pathology*. 2021;156(4):521-528. doi:https://doi.org/10.1093/ajcp/aqab014
29. Linder N, Taylor JC, Colling R, et al. Deep learning for detecting tumour-infiltrating lymphocytes in testicular germ cell tumours. *J Clin Pathol*. Feb 2019;72(2):157-164. doi:10.1136/jclinpath-2018-205328
30. Ghosh A, Sirinukunwattana K, Khalid Alham N, et al. The Potential of Artificial Intelligence to Detect Lymphovascular Invasion in Testicular Cancer. *Cancers (Basel)*. Mar 16 2021;13(6)doi:10.3390/cancers13061325
31. Noorbakhsh J, Farahmand S, Foroughi Pour A, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun*. Dec 11 2020;11(1):6367. doi:10.1038/s41467-020-20030-5

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

310

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figures

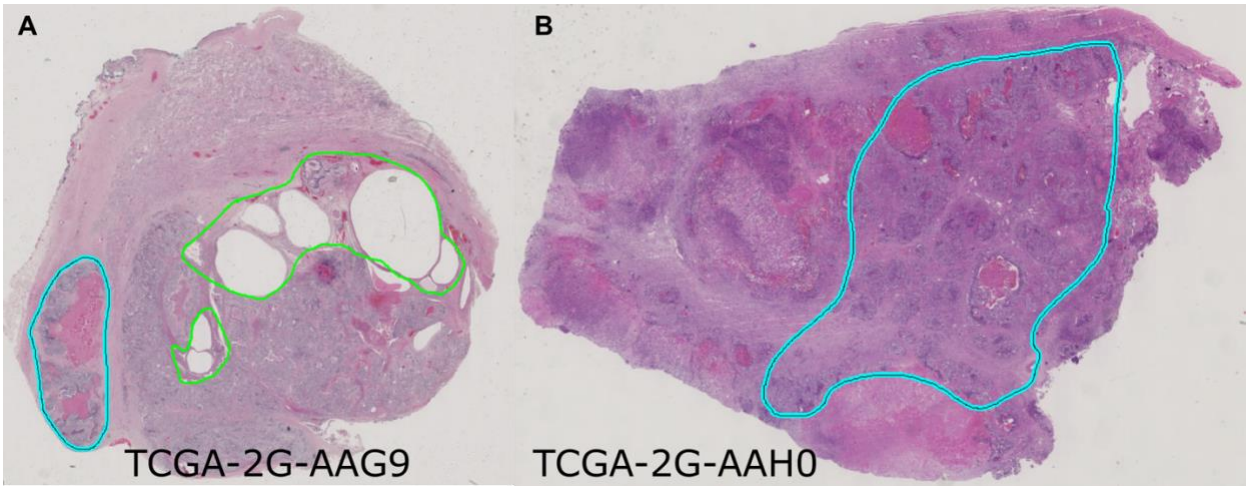
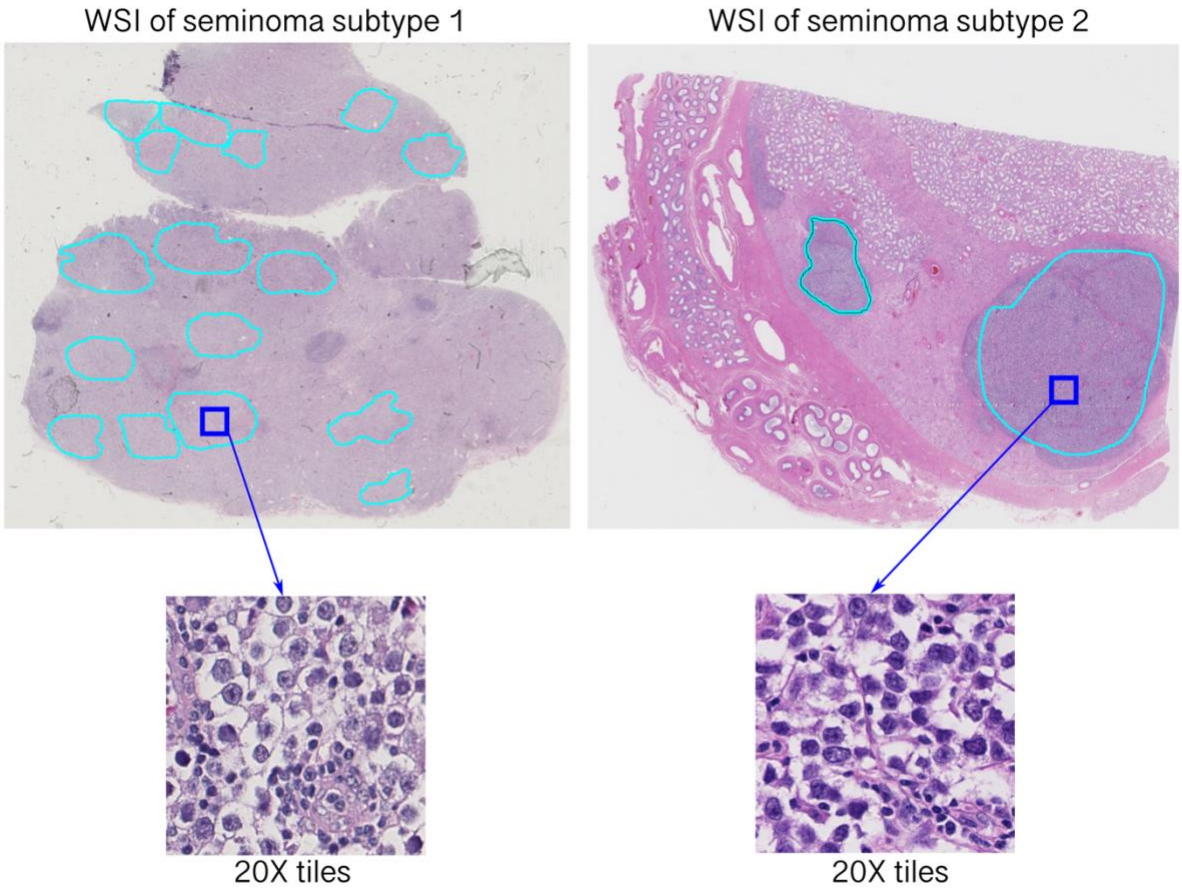


Figure 1

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

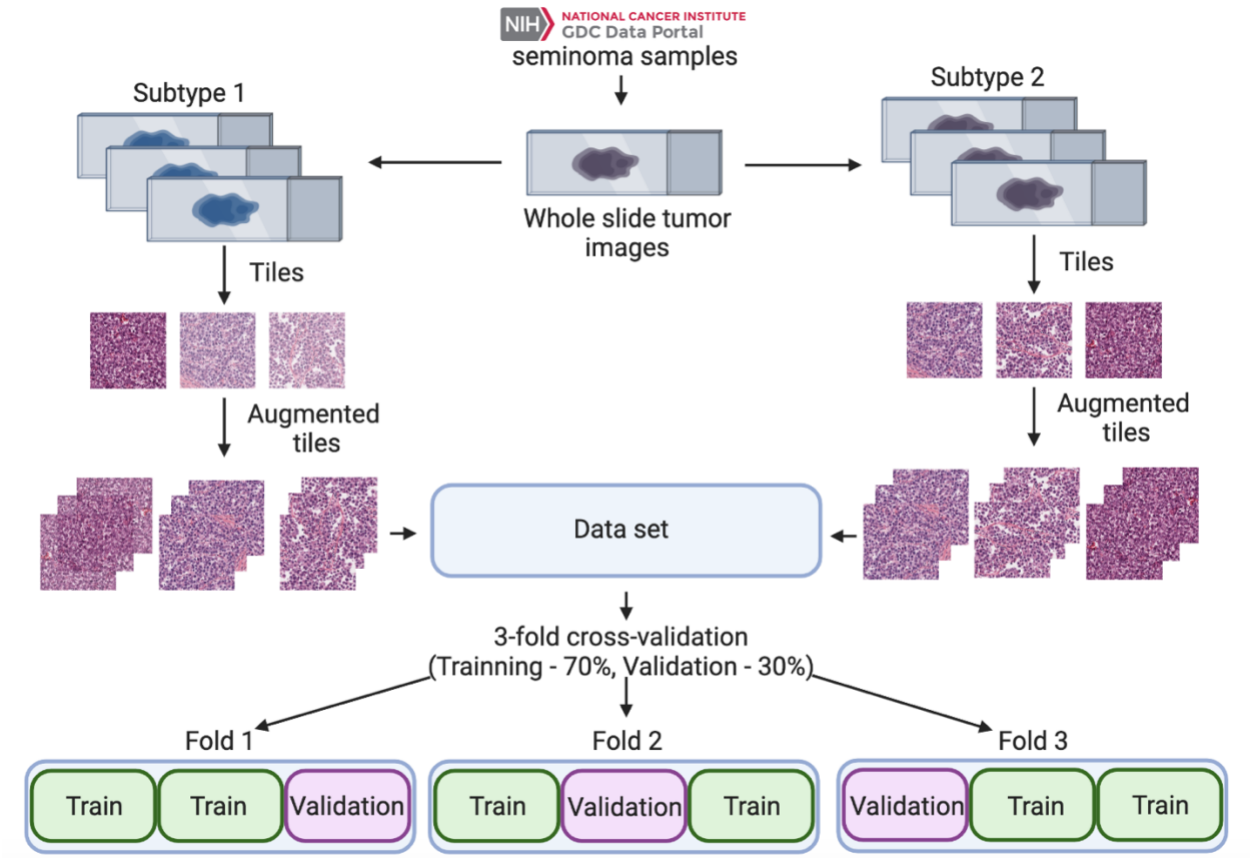


315

316 Figure 2

317

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



318

319 Figure 3

320

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

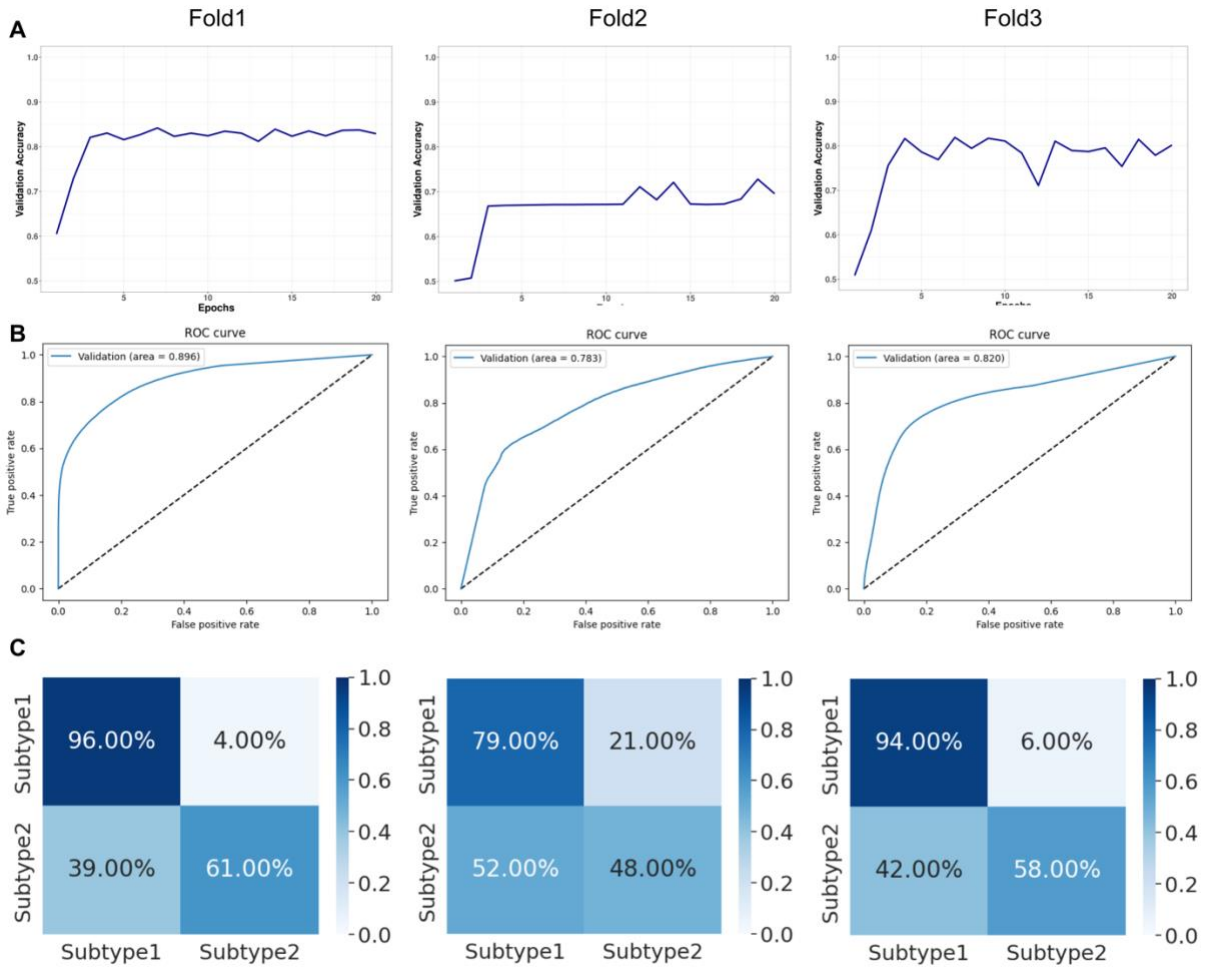
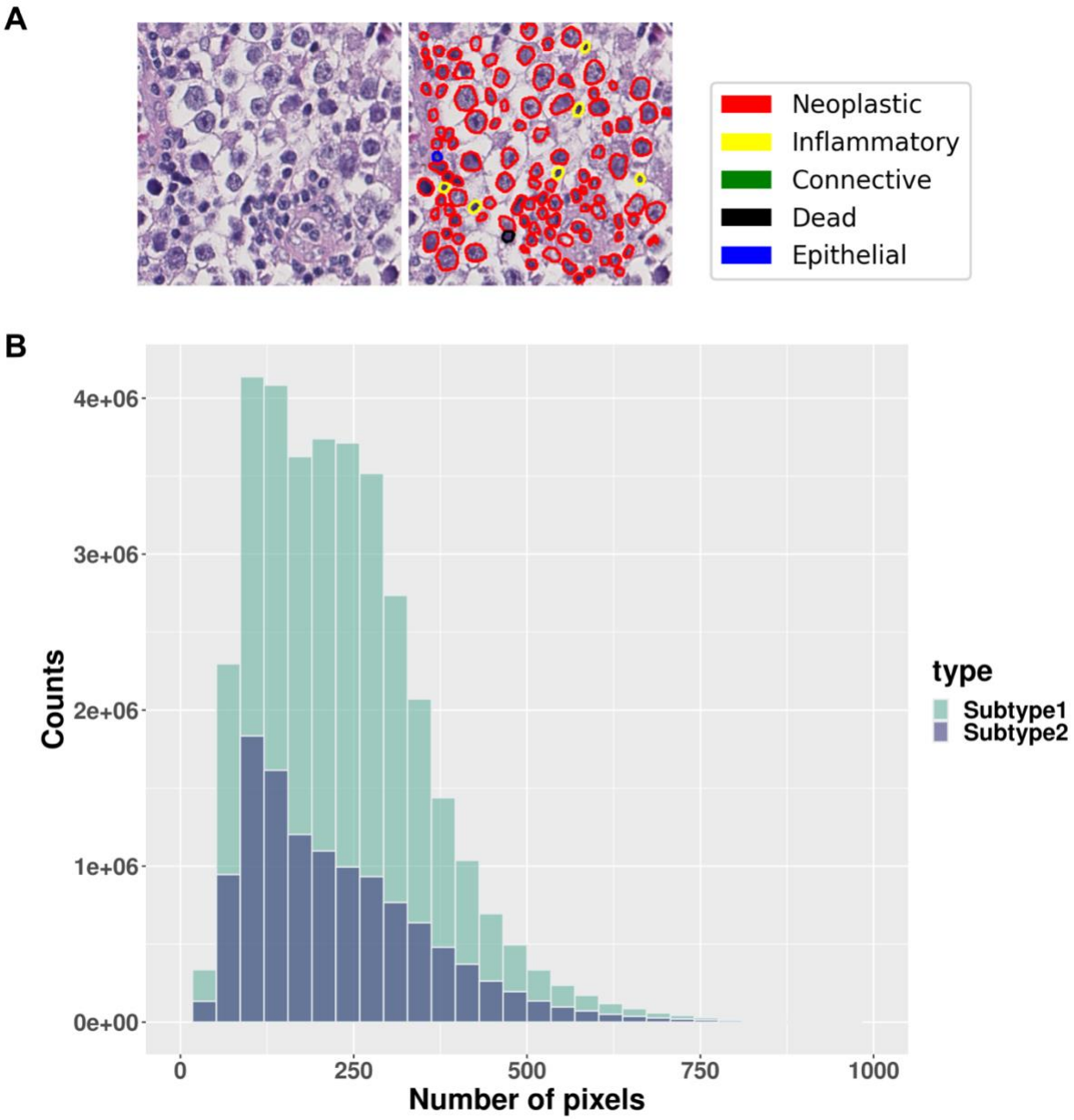


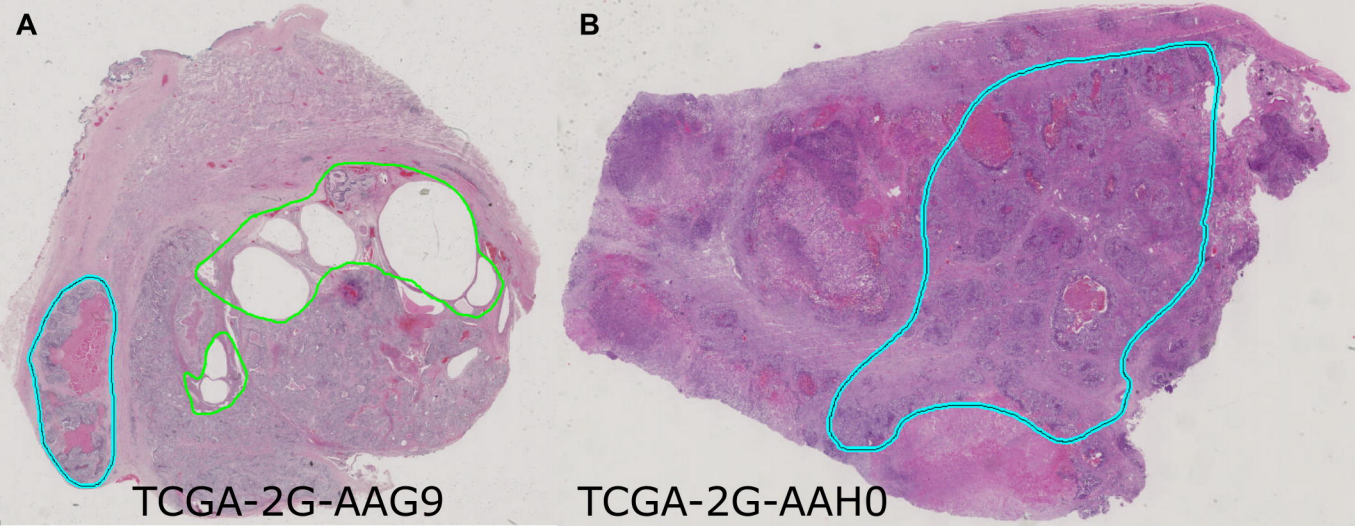
Figure 4

medRxiv preprint doi: https://doi.org/10.1101/2022.05.16.22275153; this version posted November 5, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

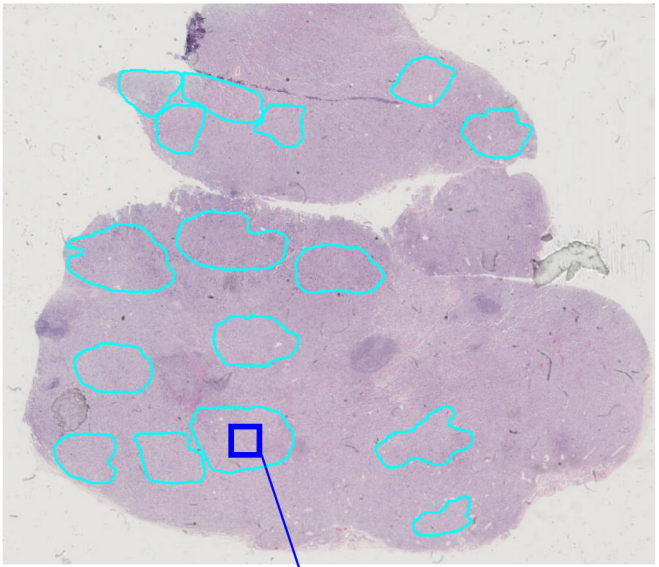


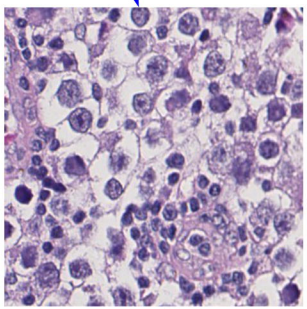
324

325 Figure 5


A
B
TCGA-2G-AAG9
TCGA-2G-AAH0

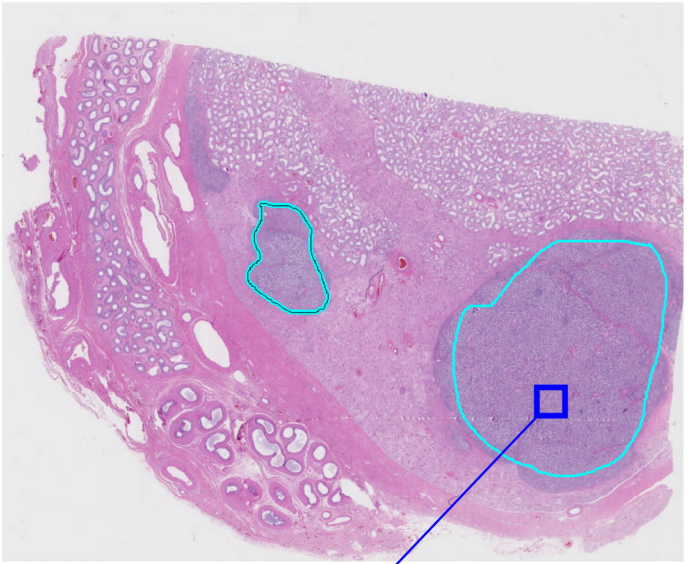
WSI of seminoma subtype 1

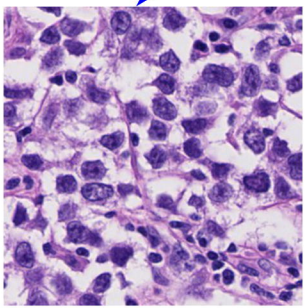




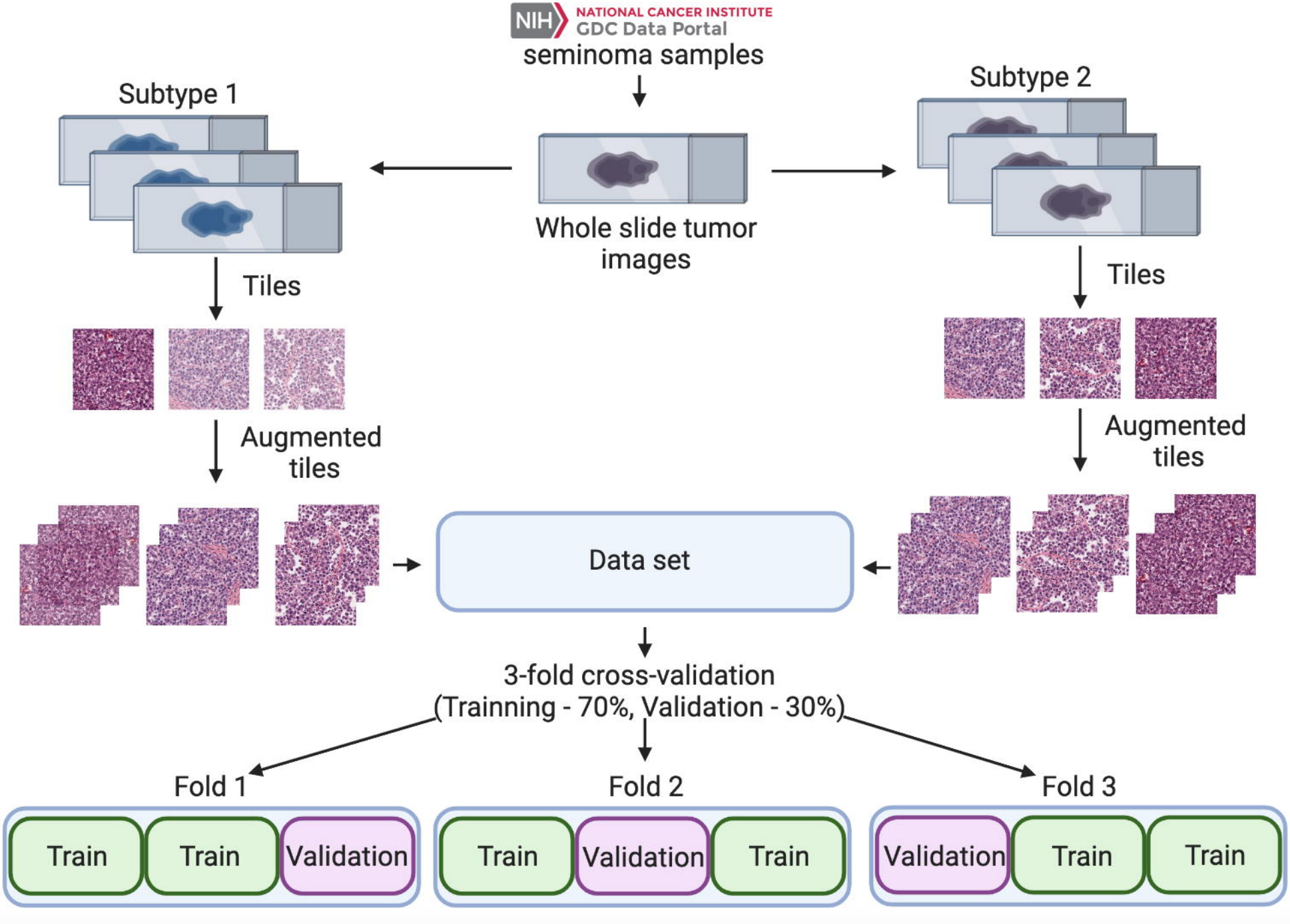
20X tiles

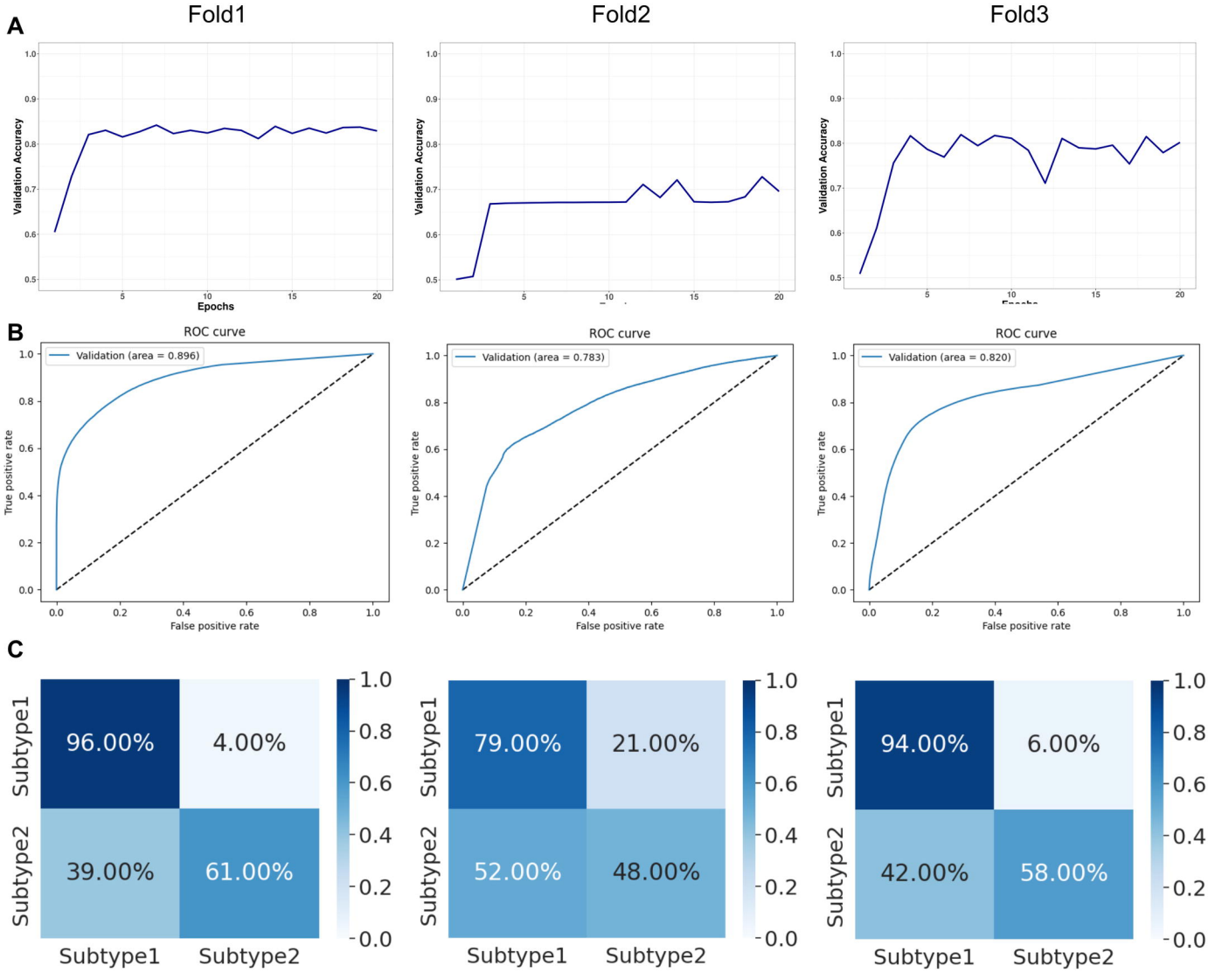
WSI of seminoma subtype 2



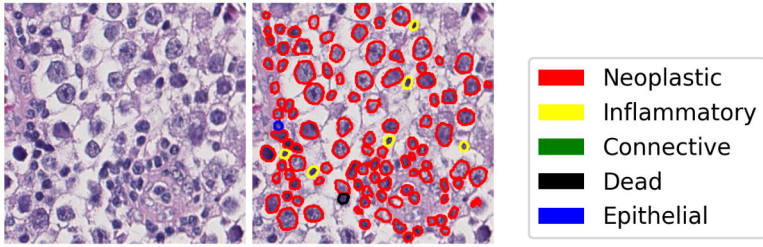


20X tiles

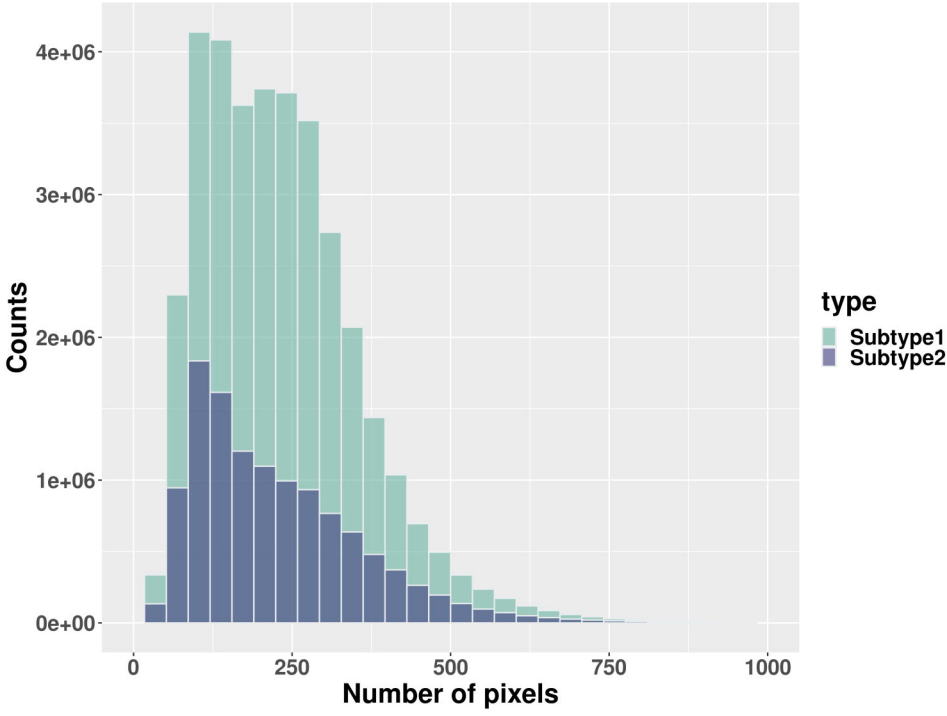

NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
seminoma samples
Subtype 1
Whole slide tumor images
Subtype 2
Tiles
Tiles
Augmented tiles
Augmented tiles
Data set
3-fold cross-validation
(Trainning - 70%, Validation - 30%)
Fold 1
Fold 2
Fold 3
Train
Train
Validation
Train
Validation
Train
Validation
Train
Train


A
Fold1
Fold2
Fold3
Validation Accuracy
Epochs
B
ROC curve
Validation (area = 0.896)
Validation (area = 0.783)
Validation (area = 0.820)
True positive rate
False positive rate
C
Subtype1
Subtype2
96.00%
4.00%
39.00%
61.00%
79.00%
21.00%
52.00%
48.00%
94.00%
6.00%
42.00%
58.00%

A


Neoplastic
Inflammatory
Connective
Dead
Epithelial

B


4e+06
3e+06
2e+06
1e+06
0e+00
Counts
0
250
500
750
1000
Number of pixels
type
Subtype1
Subtype2