Deep learning for subtypes identification of pure seminoma of the testis

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

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

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

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

* Corresponding author

E-mail: Kirill.Medvedev@UTSouthwestern.edu

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 deep learning decision making tool for 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.

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 ranks 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 secondary cancers and death [4]. Relapse occurs in approximately 20% of seminoma cases, and the underlying reasons for this phenomenon remain unclear [5]. The use of high-dose chemotherapy to treat relapsed cases can significantly exacerbate side effects. Recently we discovered two distinct subtypes of pure seminoma of the testis based on omics data [6,7]. Two identified seminoma subtypes revealed significant differences in rates of loss of heterozygosity, level of expression of lncRNA associated with cisplatin resistance, activity of double stranded DNA breaks repair mechanisms and pluripotency stage. Seminoma subtype 1 exhibits a higher pluripotency state, while subtype 2 reveals attributes of reprogramming into non-seminomatous lineages of TGCT that are more aggressive and required 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 [7]. These are features of the immune microenvironment at an early developmental stage. Moreover, subtype 2 revealed overexpression of genes related to senescence-associated secretory phenotype that might be one of the reasons for seminoma immunotherapy failure [7]. Therefore, we suggested that seminoma subtype 2 might

53 require adjustment of its treatment strategy. The development of subtype-specific therapy for
54 seminoma can reduce the risk of chemotherapy overtreatment in TGCT patients and enhance the
55 quality of life for TGCT survivors.
56 Deep learning (DL) is a powerful tool that is capable of extracting previously hidden information
57 directly from routine histopathology images of cancer tissue and can simplify, speed up and
58 automate clinical decision-making [8]. Performance of modern DL methods applied to pathological
59 data often exceeds that of human pathologists [8]. Most pathologists are working in the conditions
60 of an extreme work overload [9]. An overworked pathologist can result in misinterpretation of
61 pathological data that will affect patients' health and life quality. DL applications aim to simplify
62 and speed up routine pathological workflows and reduce pathologists' overload burden. Here, we
63 have developed a DL-based approach to examine potential histopathologic differences between
64 seminoma subtypes that were previously identified using omics data. Additionally, our goal is to
65 utilize this approach to detect and classify these subtypes based on histopathological slides. Our
66 findings demonstrate that pure seminoma subtypes cannot be classified solely based on
67 histopathological features. However, developed DL-based model revealed histopathological
68 differences between these subtypes, as indicated by the area under the ROC curve (AUC) values.
69

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 samples (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) [6]. Pure seminoma regions of interest (ROIs) were assigned and 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), reported as pure seminoma, however, that should be reported as mixed GCT, since they contain other types of GCT (teratoma and embryonal carcinoma) in addition to seminoma (Fig.1). These cases were removed from our data set.

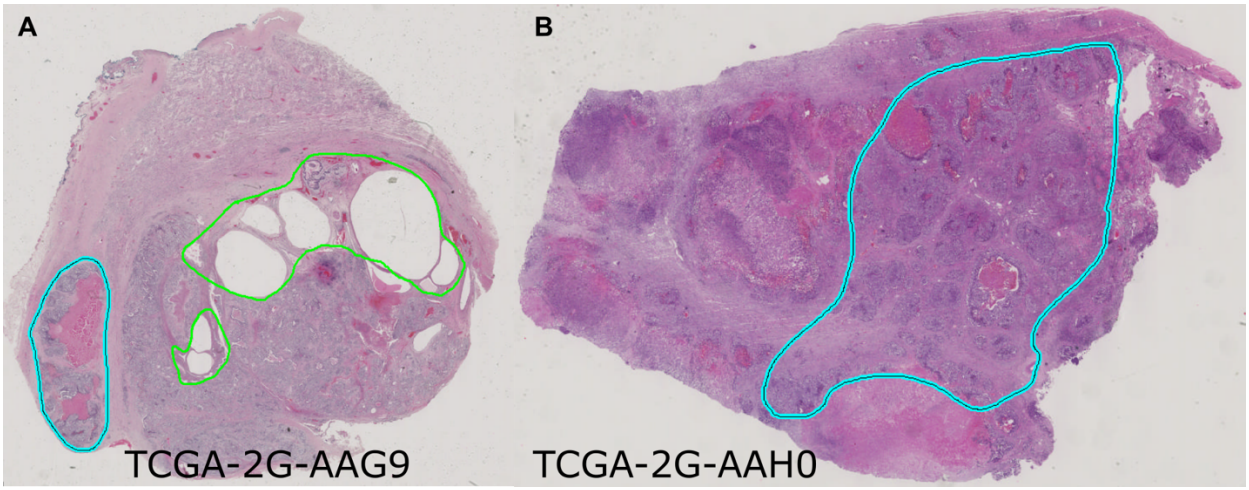


Figure 1. Slide images of samples containing additional types of GCT. (A) TCGA-2G-AAG9. **(B)** TCGA-2G-AAH0. Teratoma tissue is shown in green, embryonal carcinoma in cyan.

Verified ROIs were subsequently split into smaller tiles (300x300 pixels) at 20X magnification and with 50% overlap using DeepPath package [10] (Fig. 2). Tiles containing more than 20% of

background were removed. We conducted a manual check and excluded tiles of poor quality that contained images out of focus, defects such as scratches, dirt and folded tissue. TGCTs are relatively uncommon compared to other cancers [2], therefore the availability of histopathological data is limited. Thus, image data augmentation technique was applied to the dataset of tiles to create synthetic variations of the images and to expand the training dataset. We used following augmentations from the Imgaug library [11]: random rotation by 0°, 90°, 180°, 270°, random increase and decrease of contrast, brightness and saturation.

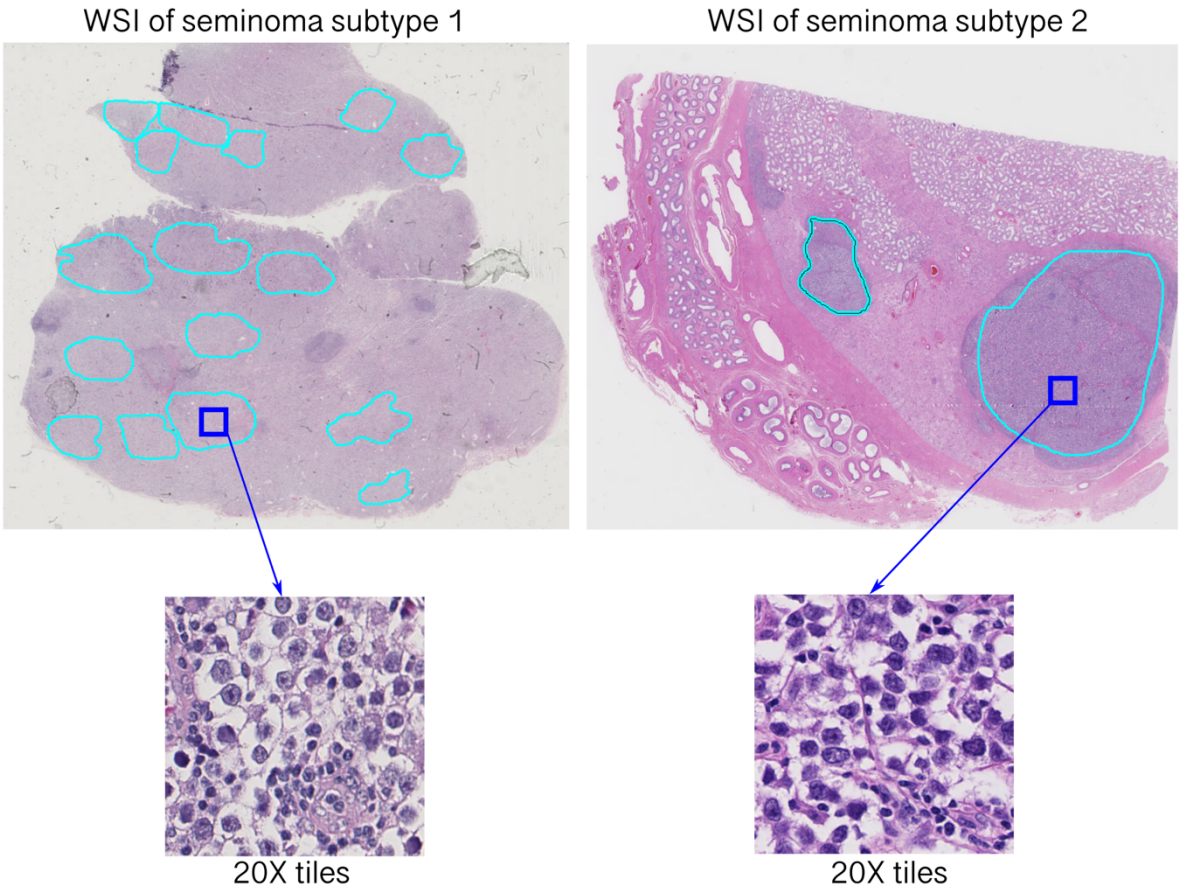


Figure 2. Extraction of tiles from annotated whole slide images (WSI) of two subtypes of pure seminoma.

Training the model

We conducted all computational experiments at BioHPC computing facility (Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, TX, USA). TensorFlow package [12] 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 [13] architecture that includes 85 convolutional layers. During the training top 29 convolutional layers were kept frozen (fixed). We used sigmoid activation and 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 [14] 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. Neural network was initialized from ImageNet-pretrained [15] weights. Models training was performed for 20 epochs with 3-fold cross-validation. Tiles belonging to 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 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.

Results and Discussion

Overview of the whole experiment is shown on Figure 3.

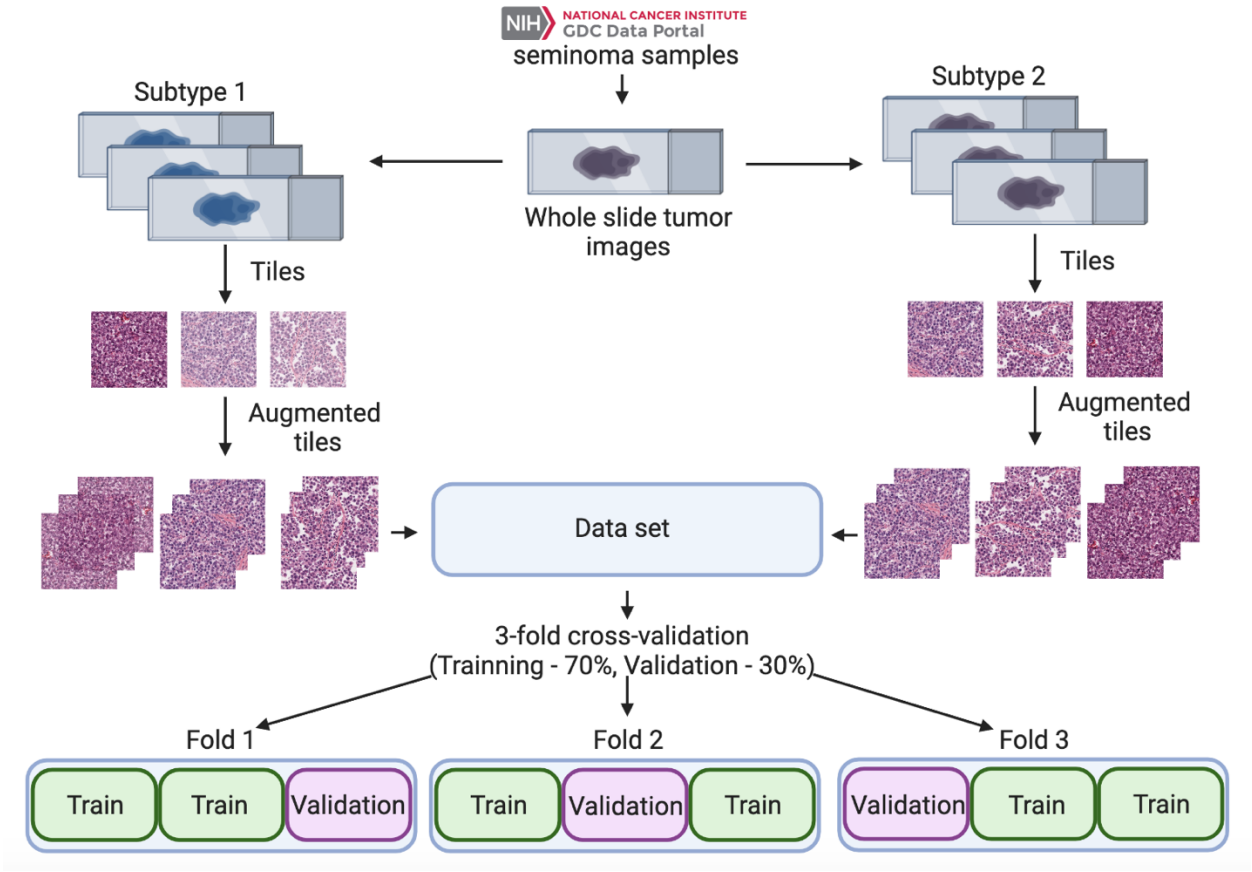


Figure 3. Overview of dataset preparation, training and validation process.

The performance of the model was evaluated using area under the ROC curve metric (Fig. 4B) and confusion matrices (Fig. 4C). 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).

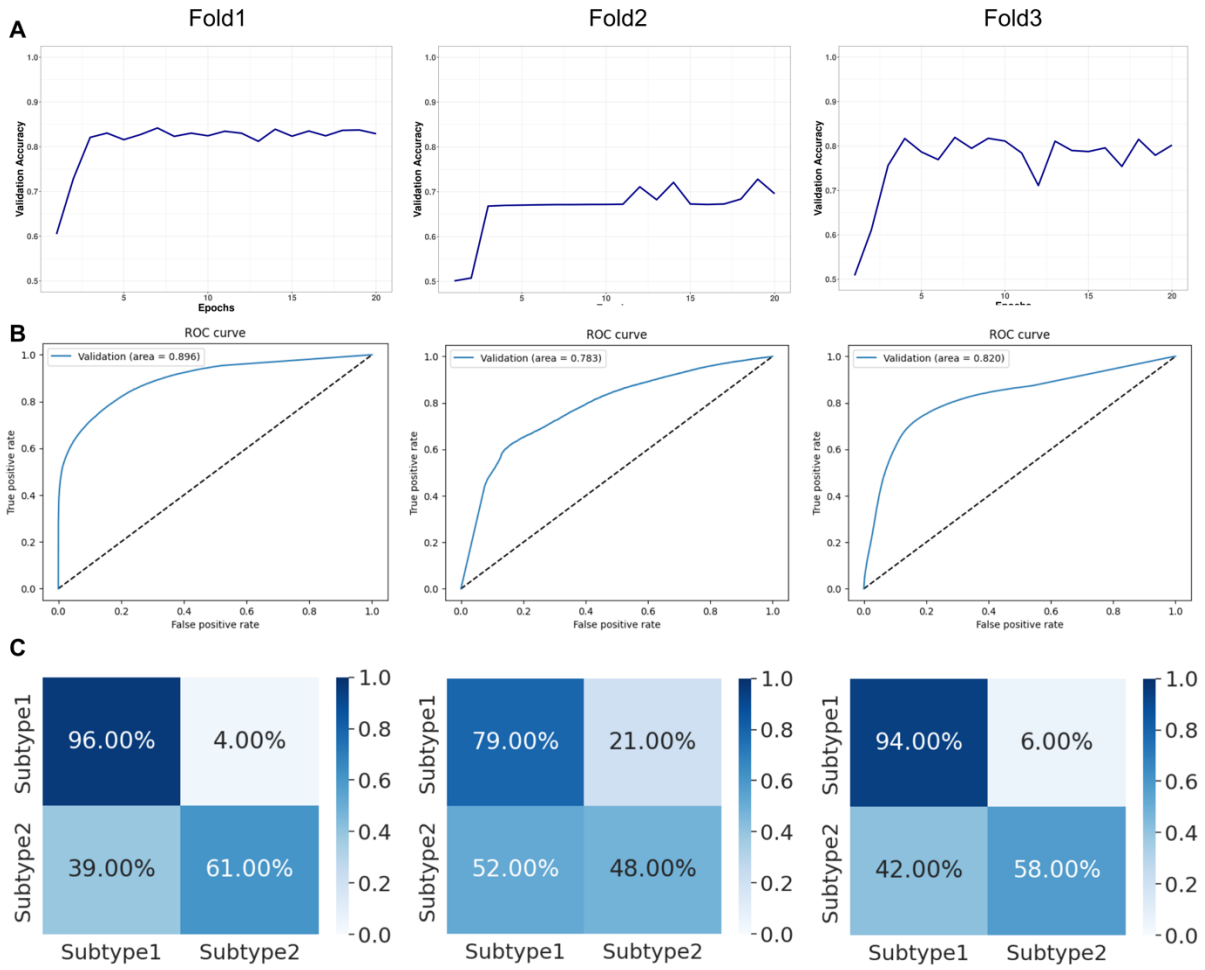


Figure 4. Validation statistics. (A) Validation accuracy. **(B)** Receiver operating characteristic (ROC) curves for validation set. **(C)** Normalized confusion matrices.

Every sample of solid tumor undergoes detailed analysis by professional pathologist that includes verification of the cancer tissue presence and annotation of tumor regions. It is the key 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% [16]. Moreover, up to one-third of clinicians do not always understand pathology reports, which causes misinterpretation and uncertainty in clinical diagnosis [17]. DL approaches applied to histopathological slide images aim

141 to significantly speed up the diagnosis and simplify their implementation into clinical workflow.
142 TGCTs and seminoma histopathology images in particular were not extensively studied using DL
143 method, and very limited studies are available nowadays. DL approached were previously applied
144 to TGCTs WSIs to detecting tumor-infiltrating lymphocytes [18], to detect lymphovascular invasion
145 [19] and to develop tumor/normal classifier [20].
146 In this report, we present our first version of the DL decision making tool for identification of pure
147 seminoma subtypes using histopathological slides. We hypothesize that consideration of
148 seminoma subtypes during development of treatment strategy will potentially improve its clinical
149 management and implementation of developed model will improve diagnosis accuracy and
150 exclude potential errors. It is especially crucial when subtypes cannot be distinguished by
151 pathologist which is the case with pure seminoma. Developed model showed capability to
152 distinguish pure seminoma subtypes (Table 1), which confirms our previous findings [6,7]. It also
153 means that morphological difference between seminoma subtypes exists. However, the correctness
154 of identification of subtype 1 samples by developed model is higher than subtype 2 samples. It is
155 possible due to certain limitations of the model. First, current model was developed using limited
156 dataset. Second, only one architecture type of CNN was tested. In future work, this model should
157 be verified using expanded dataset and several addition CNN architecture types.
158 **Table 1. Prediction of seminoma subtypes using developed DL model.** S1 – subtype 1, S2 –
159 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

In this study we developed DL-base 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.

Competing interests

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

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.).

CRediT Author Contribution

Kirill E. Medvedev: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Visualization, Writing - Original Draft, Project administration. **Liwei Jia:** Investigation, Writing - Review & Editing. **Nick V. Grishin:** Resources, Funding acquisition, Writing - Review & Editing.

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

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. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. *imgaug*. 2020. Accessed accessed 01-Feb-2022. https://github.com/aleju/imgaug
12. 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;

13. 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

14. Kingma DP BJ. Adam: A method for stochastic optimization. *arXiv*. 2014;doi:https://doi.org/10.48550/arXiv.1412.6980

15. 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;

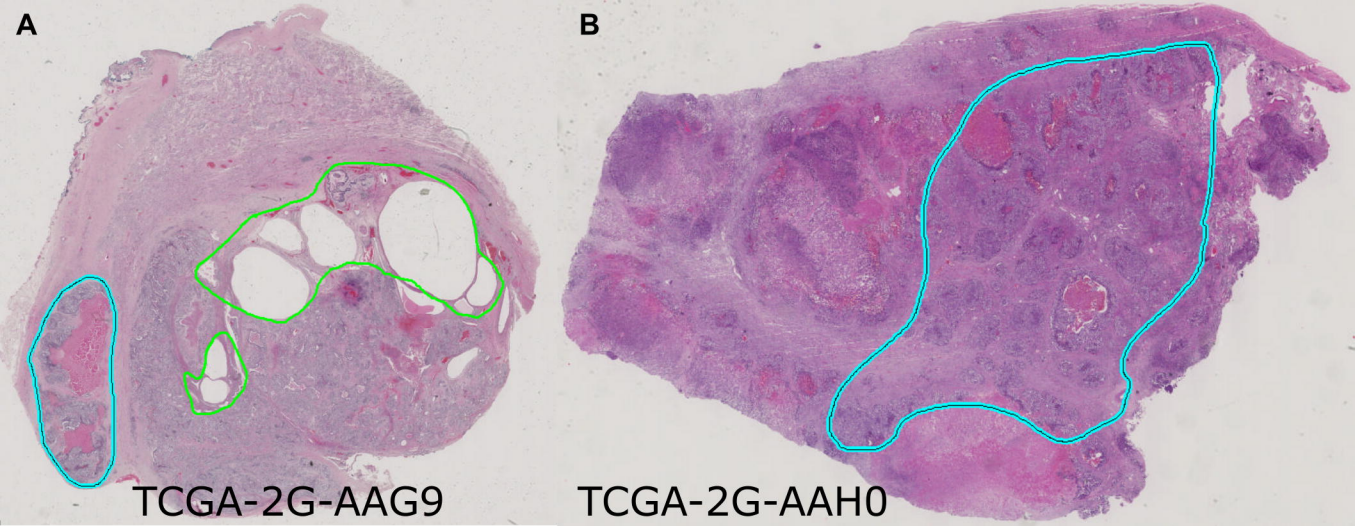
16. Ramsay AD. Errors in histopathology reporting: detection and avoidance. *Histopathology*. Jun 1999;34(6):481-90. doi:10.1046/j.1365-2559.1999.00719.x

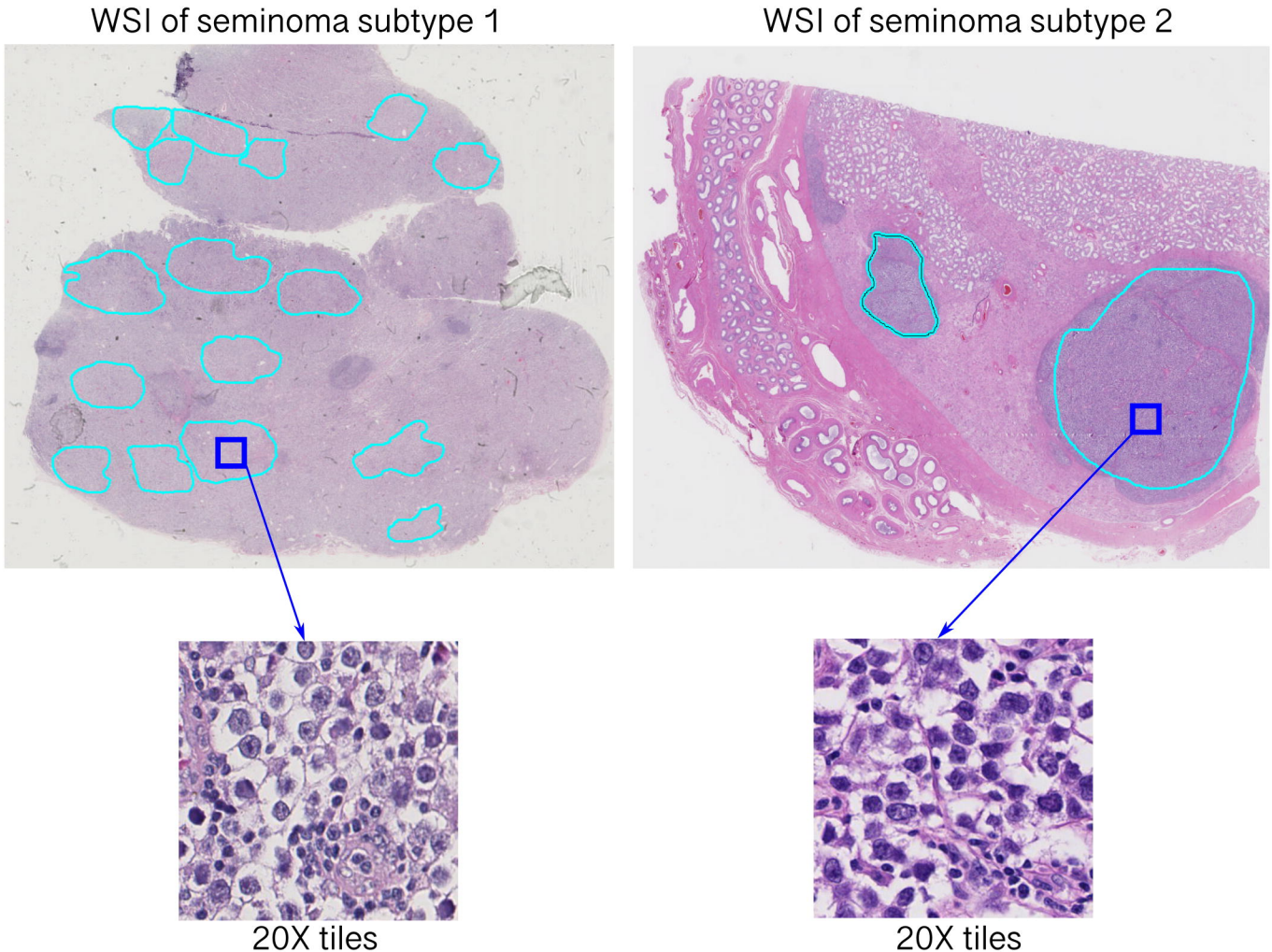
17. Mirham L HJ, Yousef GM. Addressing the diagnostic miscommunication in pathology: old challenges and innovative solutions. *American Journal of Clinical Pathology*. 2021;156(4):521-528. doi:https://doi.org/10.1093/ajcp/aqab014

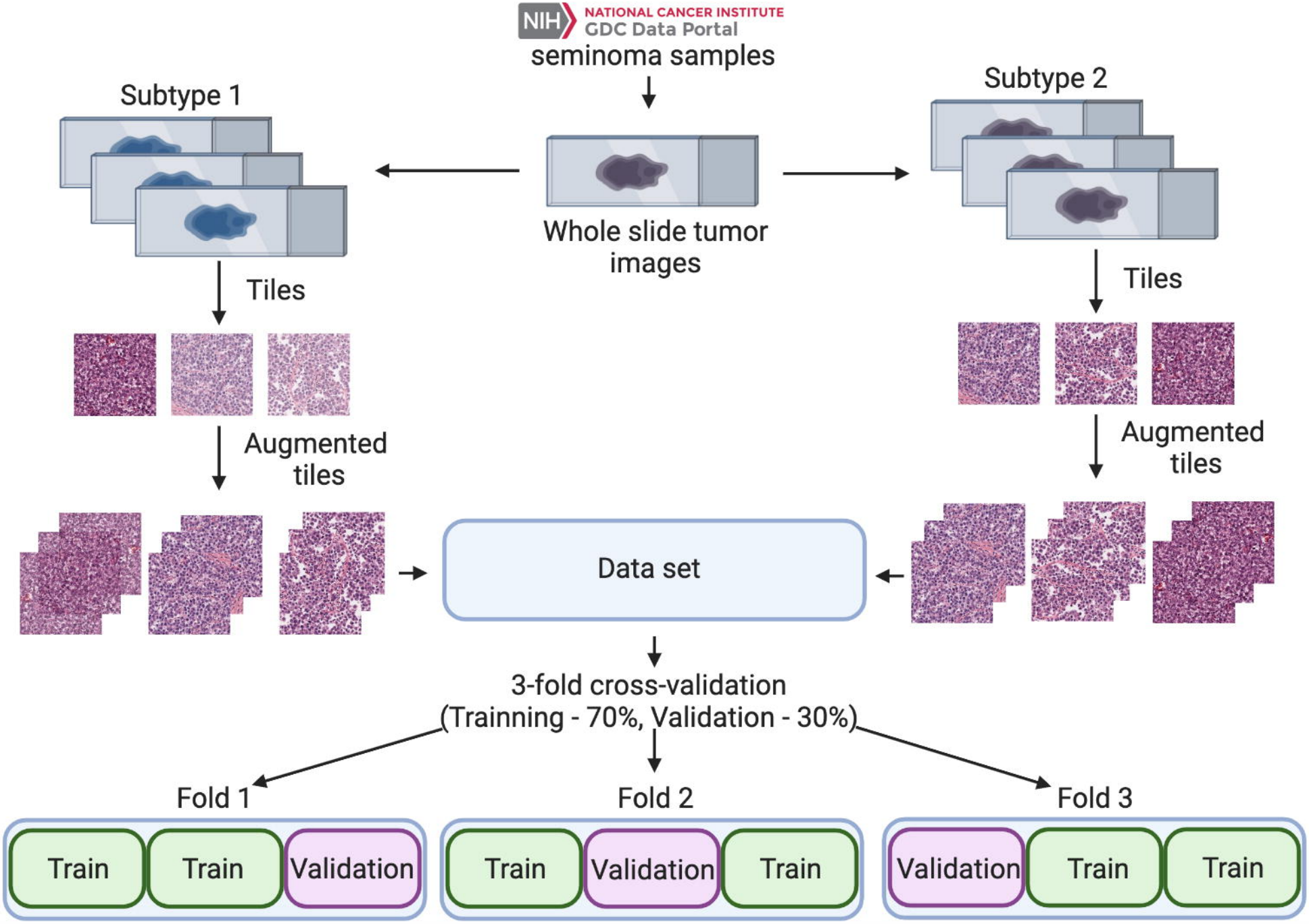
18. Linder N, Taylor JC, Colling R, et al. Deep learning for detecting tumour-infiltrating lymphocytes in testicular germ cell tumours. *J Clin Pathol*. Feb 2019;72(2):157-164. doi:10.1136/jclinpath-2018-205328

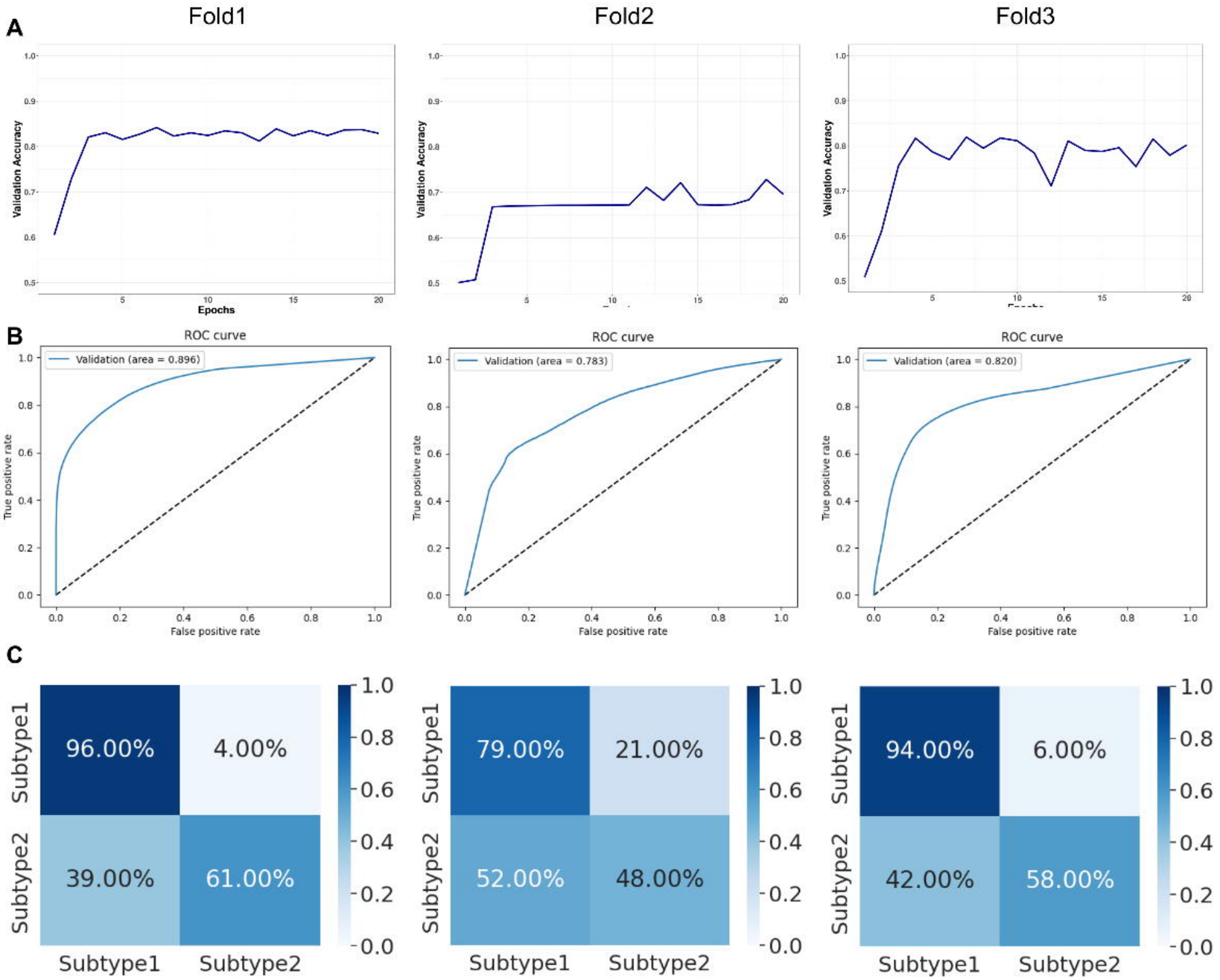
19. Ghosh A, Sirinukunwattana K, Khalid Alham N, et al. The Potential of Artificial Intelligence to Detect Lymphovascular Invasion in Testicular Cancer. *Cancers (Basel)*. Mar 16 2021;13(6)doi:10.3390/cancers13061325

20. 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


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


WSI of seminoma subtype 1
WSI of seminoma subtype 2
20X tiles
20X tiles


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
seminoma samples
Subtype 1
Subtype 2
Whole slide tumor images
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%