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

Pathological evaluation of each tumor sample is the most crucial process in the clinical diagnosis workflow. Deep learning is a powerful approach that is widely used to increase accuracy and to simplify the diagnosis process. Previously we discovered clinically relevant subtypes (1 and 2) of pure seminoma, which 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). Seminoma regions of interest (ROIs) were annotated by a genitourinary pathologist. Verified ROIs were split into tiles 300x300 pixels, which were used for model training. The model achieved the highest accuracy at the validation step with 0.933 with 0.92 area under the ROC curve.

Introduction

Seminoma is the most common histological subtype of testicular germ cell tumors (TGCT) that has the highest incidence rate within the testicular cancer group [1]. Overall, TGCT is the most common solid cancer among men aged 15-44 years [1] and the second cancer type based on calculation of life years lost per person dying of cancer, after all pediatric cancers [2]. The procedure of seminoma treatment includes orchiectomy followed by platinum-based chemotherapy using cisplatin or radiation therapy [3]. Current treatments show relatively high efficiency and high patient survival rate; however, they also cause around 40 severe adverse long-term side effects including secondary cancers, and death [4]. About 20% of seminoma cases will relapse, and the reason for this phenomenon is unclear [5]. Treatment of relapsed cases using high-dose chemotherapy will aggravate side effects drastically. Recently we discovered two distinct subtypes of pure seminoma of the testis based on omics data [6]. Two identified seminoma subtypes revealed significant differences in rates of loss of heterozygosity, DNA methylation and telomere elongation, level of expression of lncRNA associated with cisplatin resistance, activity of double stranded DNA breaks repair mechanisms and pluripotency stage. We discovered that subtype 2 revealed several molecular features that can be associated with its undergoing differentiation into non-seminomatous lineages of TGCT are more aggressive and required higher dose of chemotherapy drugs [3]. Therefore, we suggested that seminoma subtype 2 might require adjustment of its treatment strategy. We hypothesized that administration of subtype-specific therapy will significantly help to reduce high relapse rate for pure seminoma patients.

Deep learning (DL) is a powerful tool that is capable of extracting previously hidden information directly from routine histopathology images of cancer tissue and can simplify, speed up and

automate clinical decision-making [7]. Performance of modern DL methods applied to pathological data often exceed performance of pathology specialist [7]. Most of pathologist are working in the conditions of an extreme work overload [8]. An overworked pathologist can result in misinterpretation of pathological data that will affect patients' health and life quality. DL applications aim to simplify and speed up routine pathological workflows and reduce pathologists' overload burden. Here we applied DL approach to our recent results to develop decision making tool for identification of pure seminoma subtypes using histopathological slides.

Materials and Methods

Data set preparation

We used all histopathological slides available at The Cancer Genome Atlas (TCGA) data portal for 64 pure seminoma samples (156 slides). Pure seminoma regions of interest (ROIs) were assigned and verified for each slide 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.

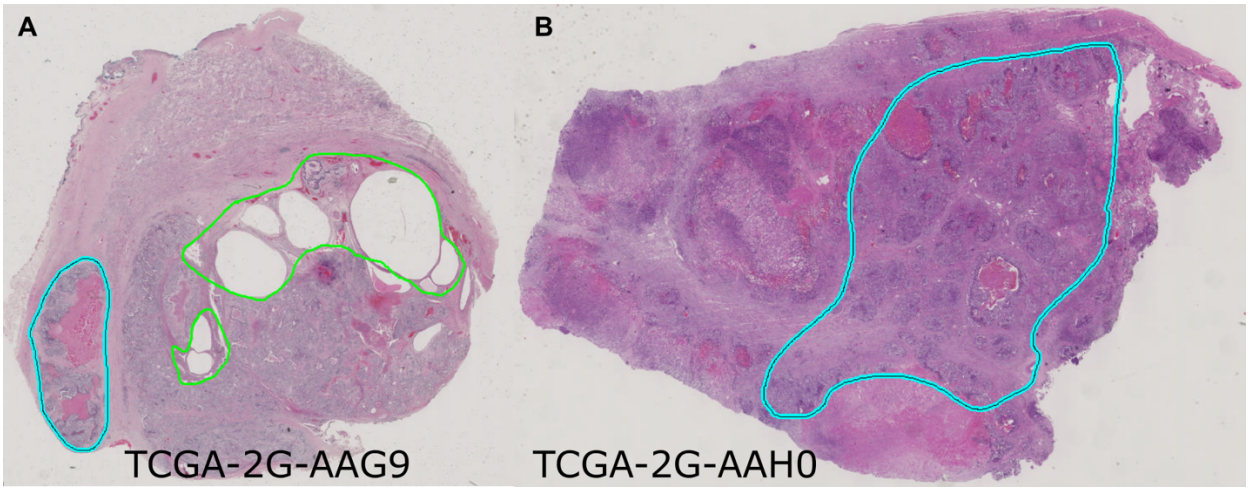


Figure 1. Slide images of samples containing additional types of GCT. (A) TCGA-2G-AAG9. **(B)** TCGA-2G-AAH0. Teratoma tissue is shown in green, embryonal carcinoma in cyan.

Verified ROIs were subsequently split into smaller tiles (300x300 pixels) at 20X magnification and with 50% overlap using DeepPath package [9]. Tiles containing more than 20% of background were removed. Overall, we obtained about 800,000 tiles which were used for model training. Image data augmentation technique was applied to the tiles dataset to create synthetic

variations of the images and to expand the training dataset (Fig.2). We used following augmentations from the Imgaug library [10]: random rotation by 0°, 90°, 180°, 270°, random increase and decrease of contrast, brightness and saturation. After augmentation procedures the total number of tiles reached about 20 million.

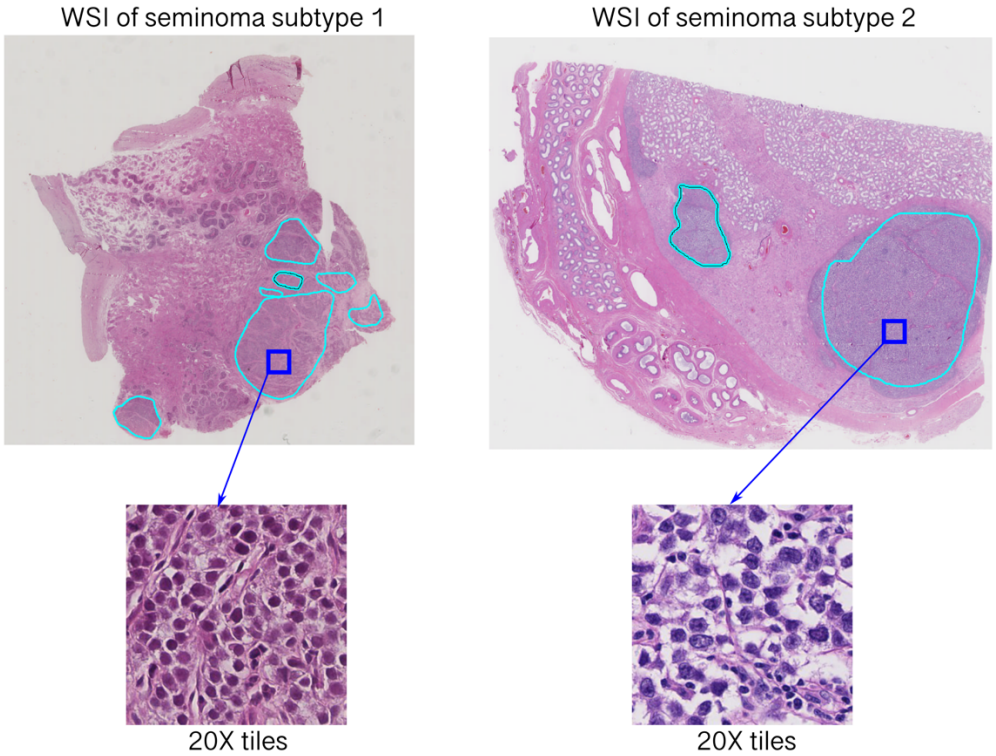


Figure 2. Extraction of tiles from annotated whole slide images (WSI) of two subtypes of pure seminoma.

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 [11] 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. Since we have binary classification goal (seminoma subtypes 1 and 2), we used Binary Cross Entropy loss function, which is used when there are only two label classes. We used an Adam optimizer [12] with a learning rate of 0.001. We employed a convolutional neural network with Inception-v3 [13] architecture with additional dense layer and sigmoid activation. This architecture has been widely used for analysis of histopathological images using DL approaches [14]. Neural network was initialized from ImageNet-pretrained [15] weights. Models training was performed for 40 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 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

Overview of the whole experiment is shown on Figure 3.

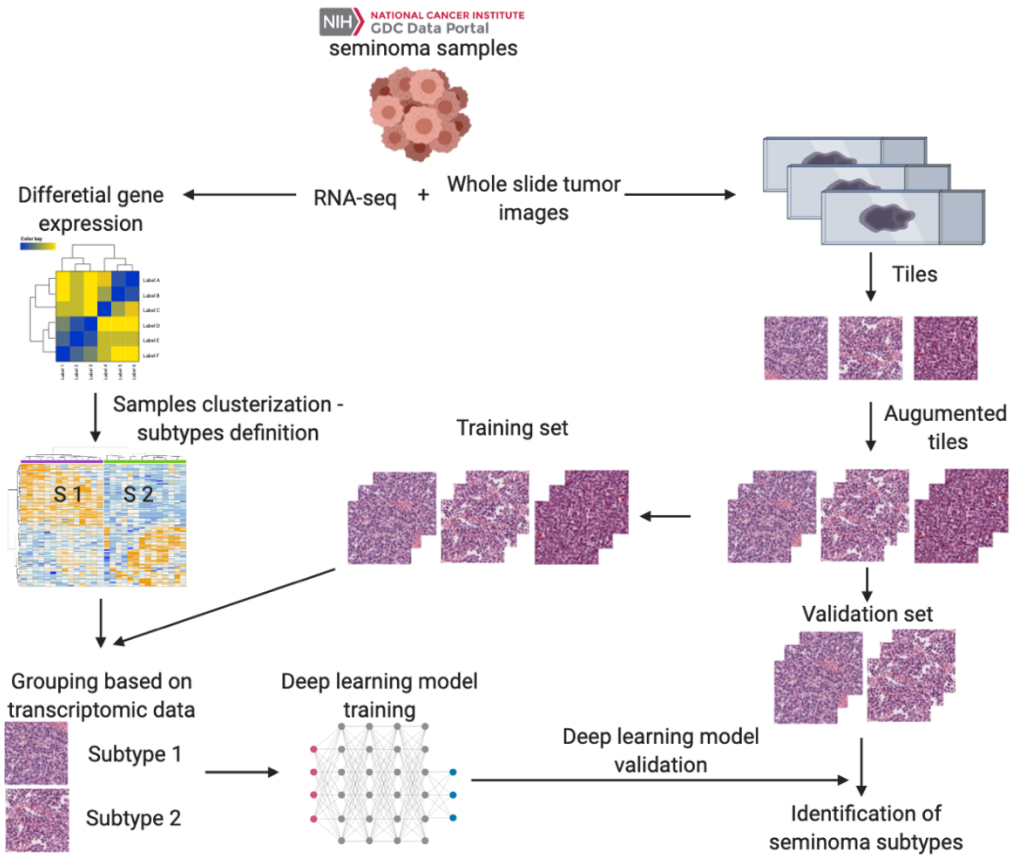


Figure 3. Overview of dataset preparation, training and validation process.

At the training stage, the model achieved the highest accuracy with 0.998. Highest achieved validation accuracy was 0.933. We also evaluate our model using area under the ROC curve metric. Our model showed AUC = 0.92 for validation set (Figure 4). Trained model for identification of pure seminoma subtypes is available in open access at GitHub (https://github.com/kirmedvedev/seminoma-subtypes).

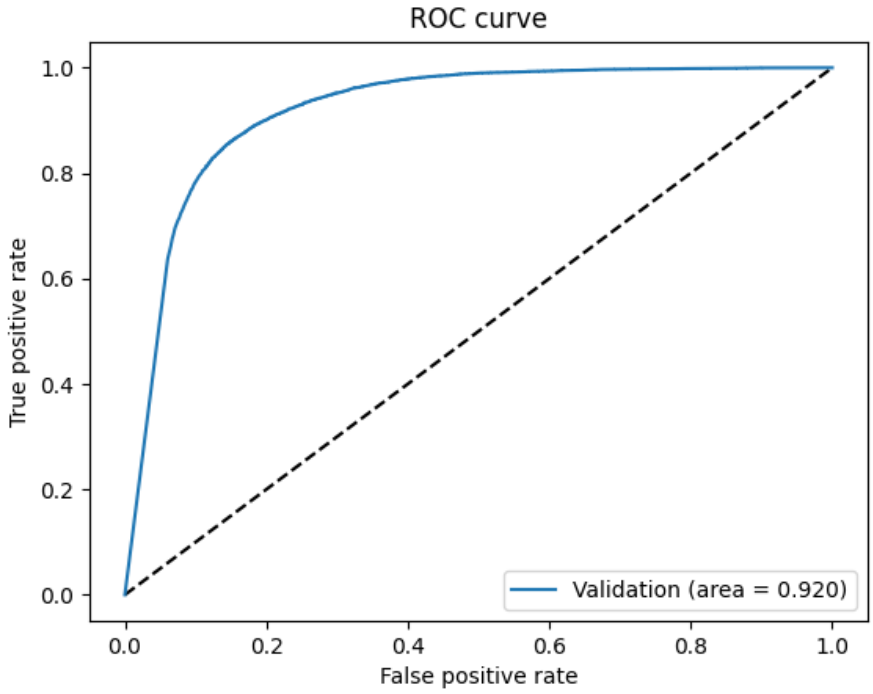


Figure 4. Receiver operating characteristic (ROC) curve for validation set.

Discussion

Every sample of solid tumor undergoes detailed analysis by professional pathologist that includes verification of the cancer tissue presence and annotation of tumor regions. This 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 to significantly speed up the diagnosis and simplify their implementation into clinical workflow. DL approached were previously applied to TGCTs WSIs to detecting tumor-infiltrating lymphocytes [18], to detect lymphovascular invasion [19].

However, to the best of our knowledge, seminomas and TGCTs histology images have not been extensively studied using DL methods.

In this report, we present our first version of the DL decision making tool for identification of pure seminoma subtypes using histopathological slides. We hypothesize that consideration of seminoma subtypes during development of treatment strategy will potentially improve its clinical management and implementation of developed model will improve diagnosis accuracy and exclude potential errors. Our model achieved the highest validation accuracy with 0.933, which is comparable with accuracy of other DL binary classifiers of histopathological images [20]. Nevertheless, there are certain limitations of developed model. First, current model was developed using limited dataset. Second, only one architecture type of CNN was tested. In future work, this model should be verified using expanded dataset and several addition CNN architecture types.

Competing interests

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

Funding

The study was supported by the grants (to N.V.G.) from the National Institutes of Health GM127390 and the Welch Foundation I-1505.

CRediT Author Contribution

Kirill E. Medvedev: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Visualization, Writing - Original Draft, Project administration.

153 **Liwei Jia:** Investigation, Writing - Review & Editing. **Nick V. Grishin:** Resources, Funding
154 acquisition, Writing - Review & Editing.
155

156 ## Acknowledgements

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

161

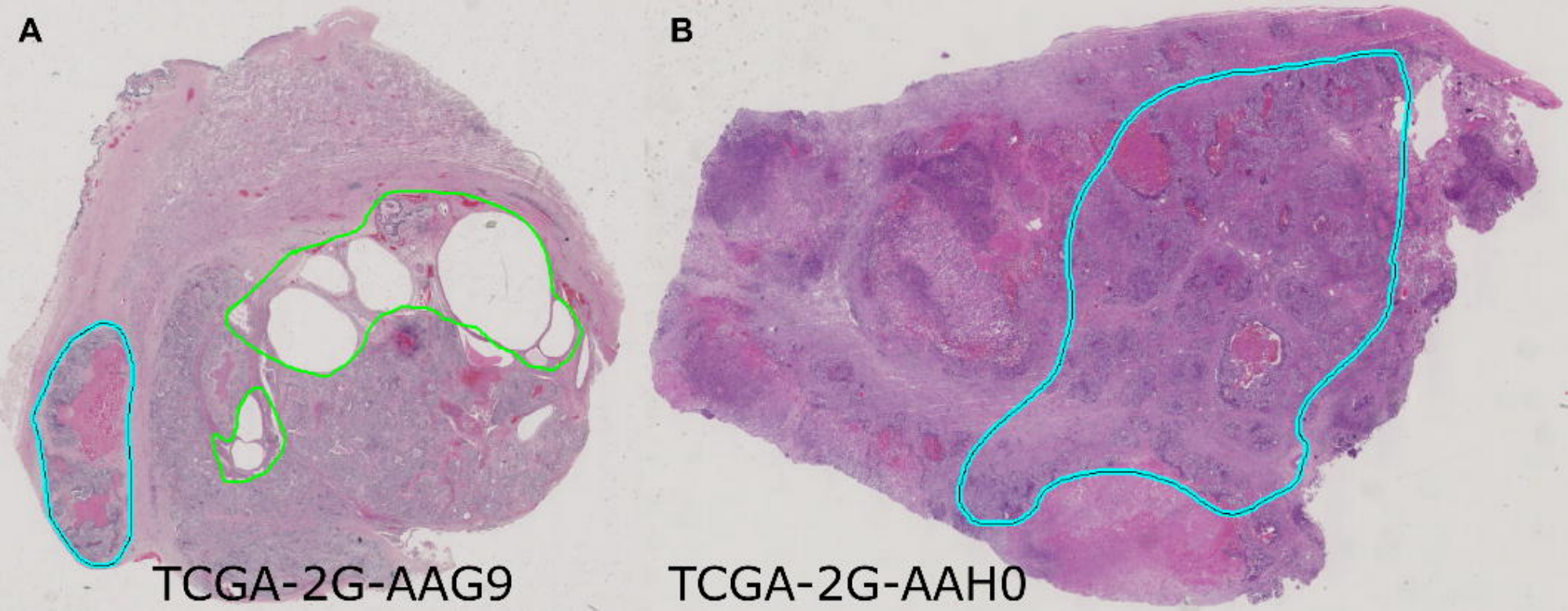
162 ## References

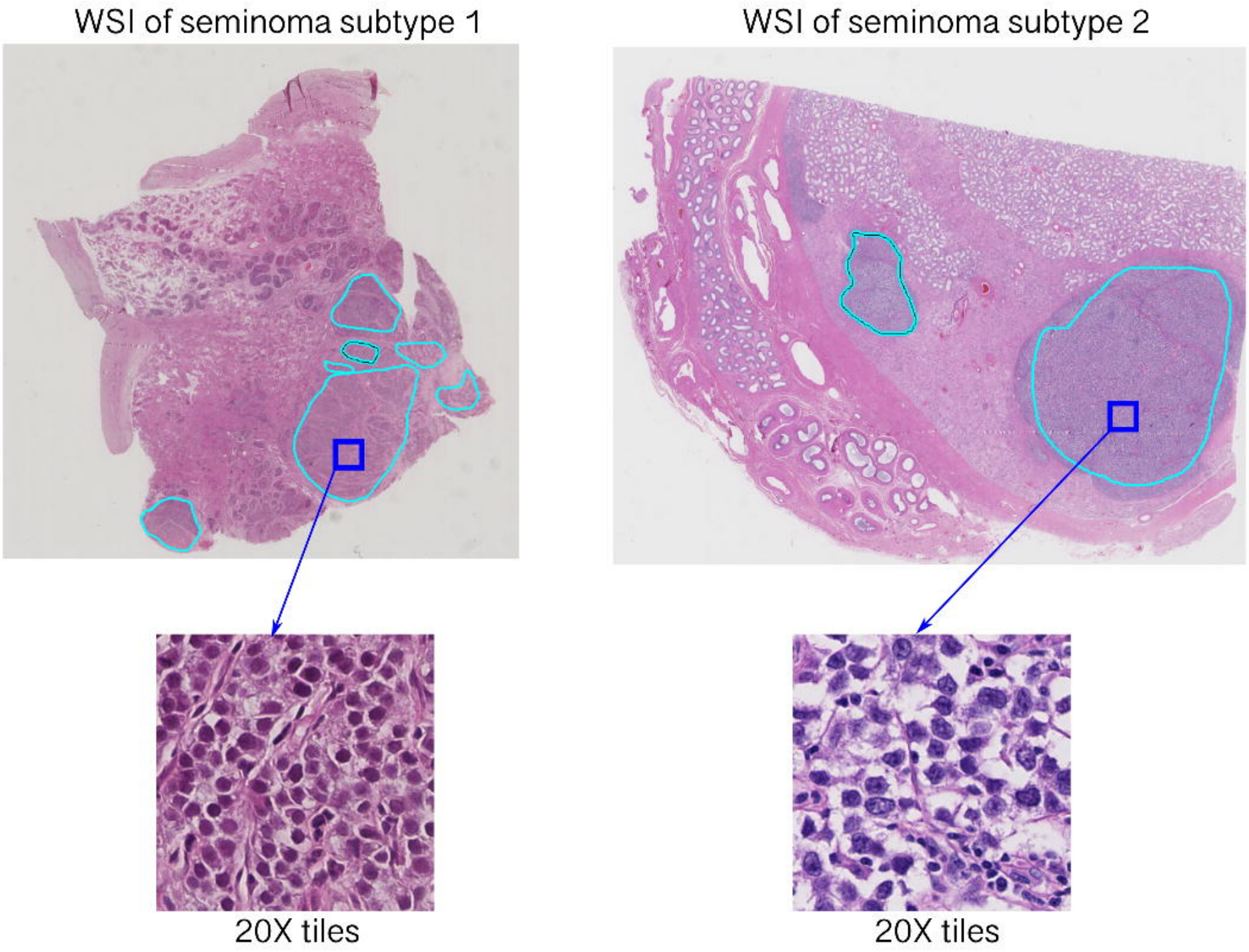
163 1. Ghazarian AA, Trabert B, Devesa SS, *et al.* Recent trends in the incidence of testicular germ
164 cell tumors in the United States. *Andrology* 2015;3:13-8.
165 2. Howlader N, Noone AM, Krapcho M, *et al*. SEER Cancer Statistics Review, 1975-2017,
166 National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2017/, based on
167 November 2019 SEER data submission, posted to the SEER web site, April 2020.
168 3. Shin YS, Kim HJ. Current management of testicular cancer. *Korean J Urol* 2013;54:2-10.
169 4. Chovanec M, Lauritsen J, Bandak M, *et al*. Late adverse effects and quality of life in
170 survivors of testicular germ cell tumour. *Nat Rev Urol* 2021;18:227-45.
171 5. Warde P, Specht L, Horwich A, *et al*. Prognostic factors for relapse in stage I seminoma
172 managed by surveillance: a pooled analysis. *J Clin Oncol* 2002;20:4448-52.
173 6. Medvedev KE, Savelyeva AV, Bagrodia A, *et al*. Integrated molecular analysis reveals two
174 distinct subtypes of pure seminoma of the testis. *bioRxiv* 2022.

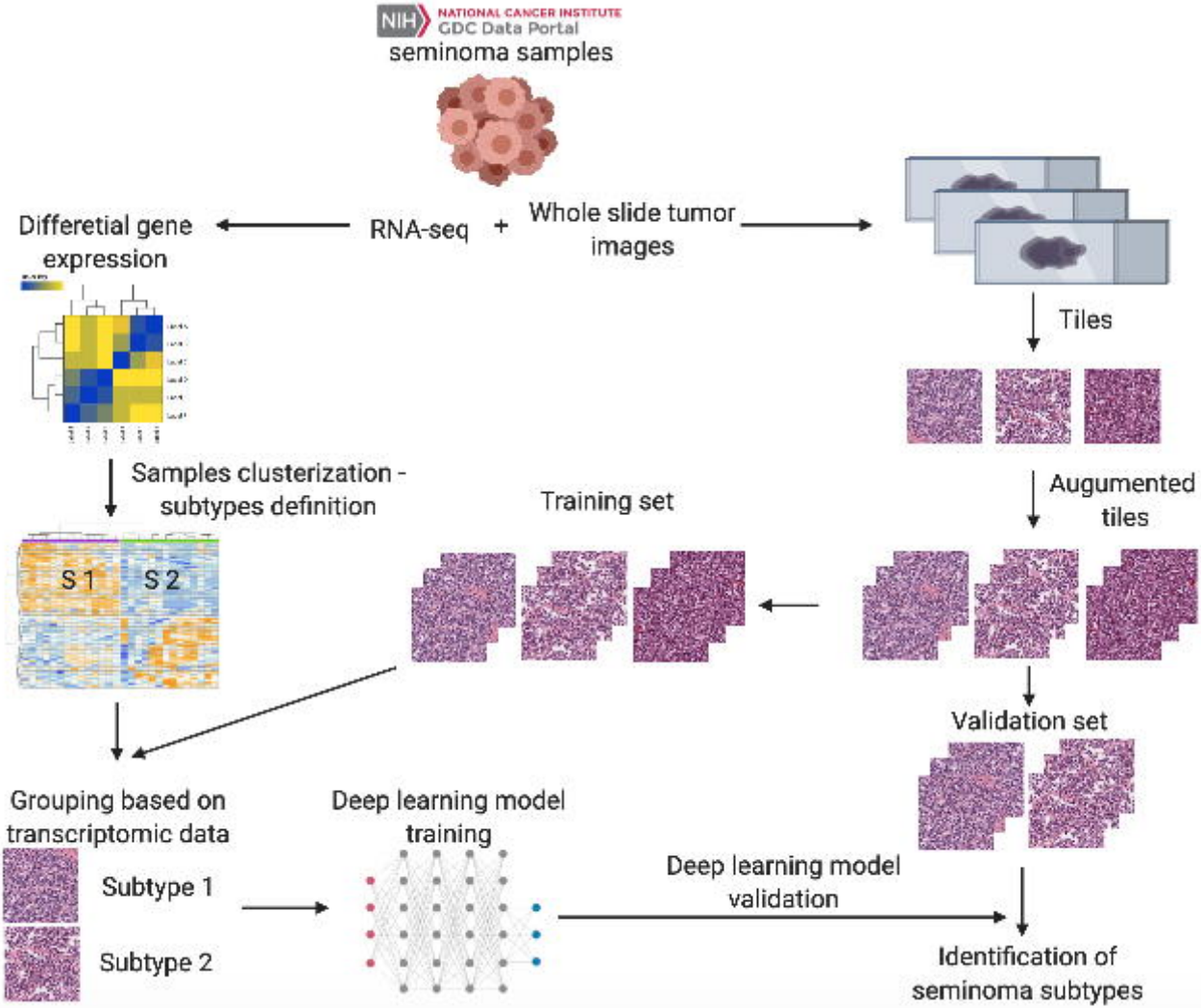
175 7. Echle A, Rindtorff NT, Brinker TJ, *et al*. Deep learning in cancer pathology: a new generation
176 of clinical biomarkers. *Br J Cancer* 2021;124:686-96.
177 8. Fritzsche FR, Ramach C, Soldini D, *et al*. Occupational health risks of pathologists-results
178 from a nationwide online questionnaire in Switzerland. *BMC Public Health* 2012;12:1-2.
179 9. Coudray N, Ocampo PS, Sakellaropoulos T, *et al*. Classification and mutation prediction from
180 non–small cell lung cancer histopathology images using deep learning. *Nat Med*
181 2018;24:1559-67.
182 10. Python library for image augmentation in machine learning experiments
183 https://github.com/aleju/imgaug
184 11. Abadi M, Barham P, Chen J, *et al*. Tensorflow: A system for large-scale machine
185 learning. In12th USENIX symposium on operating systems design and implementation
186 (OSDI 16) 2016 (pp. 265-283).
187 12. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint*
188 arXiv:1412.6980. 2014.
189 13. Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for
190 computer vision. InProceedings of the IEEE conference on computer vision and pattern
191 recognition 2016 (pp. 2818-2826).
192 14. Kather JN, Pearson AT, Halama N, *et al*. Deep learning can predict microsatellite
193 instability directly from histology in gastrointestinal cancer. *Nat Med* 2019;25:1054-6.
194 15. Deng J, Dong W, Socher R, *et al*. Imagenet: A large-scale hierarchical image database.
195 In2009 IEEE conference on computer vision and pattern recognition 2009 (pp. 248-255). Ieee.
196 16. Ramsay AD. Errors in histopathology reporting: detection and avoidance. *Histopathology*
197 1999;34:481-90.

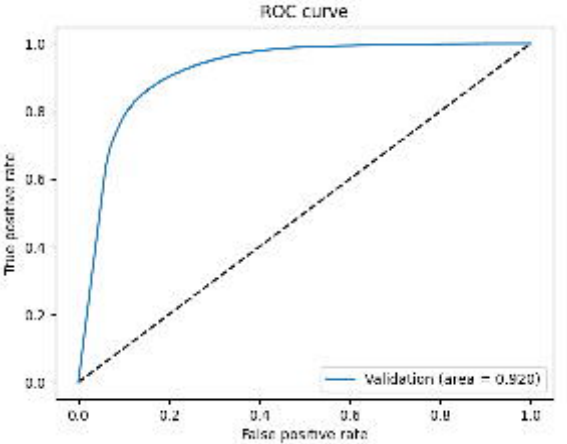
198 17. Mirham L, Hanna J, Yousef GM. Addressing the Diagnostic Miscommunication in
199 Pathology: Old Challenges and Innovative Solutions. *Am J Clin Pathol* 2021;156:521-8.
200 18. Linder N, Taylor JC, Colling R, *et al*. Deep learning for detecting tumour-infiltrating
201 lymphocytes in testicular germ cell tumours. *J Clin Pathol* 2019;72:157-64.
202 19. Ghosh A, Sirinukunwattana K, Khalid Alham N, *et al*. The Potential of Artificial
203 Intelligence to Detect Lymphovascular Invasion in Testicular Cancer. *Cancers (Basel)*
204 2021;13:1325.
205 20. Xu Y, Jiang L, Huang S, *et al*. Dual resolution deep learning network with self-attention
206 mechanism for classification and localisation of colorectal cancer in histopathological images.
207 *J Clin Pathol* 2022.

208


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
RNA-seq
+
Whole slide tumor images
Differetial gene expression
Tiles
Samples clusterization - subtypes definition
Augumented tiles
Training set
S 1
S 2
Validation set
Grouping based on transcriptomic data
Deep learning model training
Subtype 1
Subtype 2
Deep learning model validation
Identification of seminoma subtypes


ROC curve
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Validation (area = 0.920)