Ministry of Education and Science of Ukraine

ODESSA NATIONAL ACADEMY OF FOOD TECHNOLOGIES

International Competition of
Student Scientific Works

BLACK SEA SCIENCE 2021

PROCEEDINGS

ODESSA, ONAFT 2021

Ministry of Education and Science of Ukraine

Odessa National Academy of Food Technologies

International Competition of Student Scientific Works

BLACK SEA SCIENCE 2021

Proceedings

Odessa, ONAFT 2021

UDC 001.8(063)

Recommended for print by the Academic Council of
Odessa National Academy of Food Technologies
on April 6, 2021, Protocol No. 13

Black Sea Science 2021: Proceedings of the International Competition of Student Scientific Works / Odessa National Academy of Food Technologies; B. Iegorov, M. Mardar (editors-in-chief.) [*et al.*]. – Odessa: ONAFT, 2021. – 731 p.

Proceedings of International Competition of Student Scientific Works «Black Sea Science 2021» contain the works of winners of the competition.

The author of the work is responsible for the accuracy of the information.

Organizing committee:

Prof. Bogdan Iegorov, D.Sc., Rector of Odessa National Academy of Food Technologies, Head of the Committee

Prof. Maryna Mardar, D.Sc., Vice-Rector for Scientific and Pedagogical Work and International Relations of Odessa National Academy of Food Technologies, Deputy Head of the Committee

Prof. Stefan Dragoev, D.Sc., Vice-Rector for Scientific Work and Business Partnerships of University of Food Technologies (Bulgaria)

Prof. Baurzhan Nurakhmetov, D.Sc., First Vice-Rector of Almaty Technological University (Kazakhstan)

Prof. Mircea Bernic, Dr. habil., Vice-Rector for Scientific Work of Technical University of Moldova (Moldova)

Prof. Jacek Wrobel, Dr. habil., Rector of West Pomeranian University of Technology (Poland)

Prof. Michael Zinigrad, D.Sc., Rector of Ariel University (Israel)

Dr. Mei Lehe, Ph.D., Vice-President of Ningbo Institute of Technology, Zhejiang University (China)

Prof. Plamen Kangalov, Ph.D., Vice-Rector for Academic Affairs of "Angel Kanchev" University of Ruse (Bulgaria)

Dr. Alexander Sychev, Ph.D., Assoc. Professor of Sukhoi State Technical University of Gomel (Belarus)

Dr. Hanna Lilishentseva, Ph.D., Assoc. Professor, Head of the Department of Merchandise of Foodstuff of Belarus State Economic University (Belarus)

Prof. Heinz Leuenberger, Ph.D., Professor of the Institute of Ecopreneurship of University of Applied Sciences and Arts (Switzerland)

Prof. Edward Pospiech, Dr. habil., Professor of the Institute of Meat Technology of Poznan University of Life Sciences (Poland)

Prof. Lali Elanidze, Ph.D., Professor of the Faculty of Agrarian Sciences of Iakob Gogebashvili Telavi State University (Georgia)

Dr. V. Kozhevnikova, Ph.D., Senior Lecturer of the Department of Hotel and Catering Business of Odessa National Academy of Food Technologies, Secretary of the Committee

3. INFORMATION TECHNOLOGIES, AUTOMATION AND ROBOTICS

COMPUTERIZED DIAGNOSIS OF PROSTATE CANCER BASED ON WHOLE SLIDE HISTOLOGY IMAGES AND DEEP LEARNING METHODS

Authors: Valery Malyshev
Dmitry Voynov
Elizabeth Lapo
Advisor: Vassili Kovalev
United Institute of Informatics Problems (Belarus)

Abstract. *This paper addresses the problem of the diagnosis of prostate cancer and providing the final score in the commonly used ISUP scale. The input data consisted of 10 616 whole-slide histological images with the size of the largest dimension up to 100 000 pixels and 22 089 of their image tiles of 256x256 pixels in size. Several approaches were examined including conventional methods based on feature extraction followed by suitable classifiers as well as various deep learning techniques based on convolutional neural networks of different architectures. During the experiments, two deep learning approaches have been shown to be the most efficient. The first one implements a technique of sequential, tile-by-tile image classification using ResNet50 as a feature extractor followed by a prediction of final ISUP score based on the informative features using a small convolutional network. The second approach predicts the final ISUP score using an ensemble of 4 deep learning models for regression from a subset of the most prominent image tiles first, combined into one single pseudo-image. Being independently tested on an unknown image dataset that was not available for authors, these approaches achieved the prediction accuracy of 0.8110 and 0.9277 respectively.*

Keywords: *prostate cancer, whole slide histology, deep learning, convolutional neural networks*

I. INTRODUCTION

Prostate cancer is one of the most common types of cancer among men. More than 1 million cases are diagnosed every year and more than 350 thousand people die due to this disease [1]. Early detection and correct identification of the tumor aggressiveness is a crucial step in reducing the prostate cancer mortality rate. To define prostate cancer stages and whether it is likely to grow rapidly, the Gleason grading system [2] was introduced. A Gleason score is assigned based upon a high-resolution histological image of a tissue sample obtained as a result of biopsy scanning. The existing versions of the grading system may slightly differ, however, the most common one includes the following integer values: 0, 3, 4 and 5. Score 0 corresponds to healthy tissue, score 3 is assigned to a biopsy site with minimal malignancy and with a positive prognosis of survival under appropriate treatment, score 5 denotes a maximum of

cancer aggressiveness with the worst prognosis. Accordingly, score 4 is an intermediate stage between scores 3 and 5. Considering that an image of a tissue sample can be very heterogeneous and cancerous regions might have different Gleason scores, modern medicine prefers to use a two-level grading system. The first level corresponds to a tissue site, which area is the largest one on the studied tissue sample, and the second level corresponds to the one with the second largest area. The assessed level numbers are written through the "+" sign. For example, the tissue sample is coded "3+5" when its biggest area represents the initial stage of the disease and its next largest section represents the most aggressive stage. The sample is coded "5+5" when its last stage is completely dominant. Subsequently, an aggregated score is calculated either as a sum of the indicated numbers, or, more often, as an indicator defined in the form of a lookup table: 0 + 0 = 0, 3 + 3 = 1, 3 + 4 = 2, 4 + 3 = 3, 3 + 5 = 4, 4 + 5 = 5 and 5 + 5 = 5. The integrated scale is called the ISUP scale (ISUP Score, [2]). The "nonlinearity" of the ISUP scale reflects a much greater danger of high scores, especially of the maximum Gleason value 5. This can be explained by the fact that even if a small amount of highly aggressive tissue is found, there is a very high probability of rapid development of the disease.

The discussed problem can be viewed as a machine learning problem, where a certain value from the ISUP scale range (0-5) should be predicted using an image of tissue as an input. Since ISUP is a scale of cancer aggressiveness and all possible scores are ordered, it is important not only to predict the correct answer as some abstract class number but also to have the answer as close to the true score as possible. In the first case, i.e. without taking into account the presence of the order relation of the class values, we face a classification problem. In the second case, the problem should be considered a regression.

It should also be noted that the diagnosis of prostate cancer is a rather difficult task even for experienced pathologists [1, 3]. As a result, there may be a significant discrepancy between the conclusions of several experts on the same patient. The complexity of the diagnosis depends on many factors, which include large heterogeneity of the tissue sample, the high variability of the morphological structure and the spatial-color patterns within only one Gleason score, significant differences in the histological patterns of different patients, the presence of artifacts in the preparation, staining and scanning of tissue samples and some other factors [4, 5].

II. LITERATURE ANALYSIS

During the last decade, deep learning has significantly improved performance on a variety of tasks, including diagnosis and assessment of the degree of prostate cancer. Thus, in works [5, 6], deep learning methods have shown good results in solving problems of binary classification of prostate tissue (cancer versus normal, early stage of cancer versus late-stage). Nevertheless, using software for cancer grading nor as an independent predictor (see [5, 7], Kappa 0.62 and 0.70, respectively), neither as an additional step in histopathologists' decision-making process ([8], Kappa 0.733) led to a satisfactory quality of assessment Gleason scores. The number of studied WSIs was significantly smaller in all the mentioned works compared to this study.

III. OBJECT, SUBJECT, AND METHODS OF RESEARCH

The object of this research was an application of deep learning methods in medical imaging, and the subject was a development of computational tools, algorithms and models for prostate cancer diagnosis by labeling images with ISUP grade. This work was aimed to study and develop tools for automatic analysis and classification of whole-slide histological images (WSI) experimentally, as well as to obtain generally accepted quantitative ISUP grades of the prostate cancer aggressiveness degree to be potentially used as an additional opinion by pathologists. The research was done as a part of an international competition for the diagnosis of prostate cancer PANDA [6], which more than a thousand teams from different countries have taken part in. According to the results of independent testing on a set of images that were inaccessible to the authors, this work ranked No. 18 out of 1010 [9].

WSI images can reach significant sizes, up to 100,000 pixels in each dimension with a memory capacity of up to 10 gigabytes (without compression). Such sizes make it difficult to use traditional methods, and existing deep learning methods are not suited for this purpose either. Also, parallel computation algorithms on graphics accelerators (GPUs) for deep learning do not allow placing the required number of images in graphics memory simultaneously. Therefore, it was necessary to perform a certain preprocessing to solve this problem. The common approaches are to reduce the size of the initial problem, which might lead to a loss of certain information in the process, and to divide it into several smaller subtasks that can be solved separately by existing methods. We followed both of these approaches to prepare the initial data made of prostate WSI and obtained several datasets, which were used in further method development. Details about input data and different techniques we used to prepare it for analysis are listed in section 3.1. For datasets we mainly used separation of images into separate square regions. Those regions were of small dimensions, therefore could be used for various algorithms applications without the risk of heavy memory consumption. For clarity we call these square regions of WSI images are called tiles in subsequent description of developed methods.

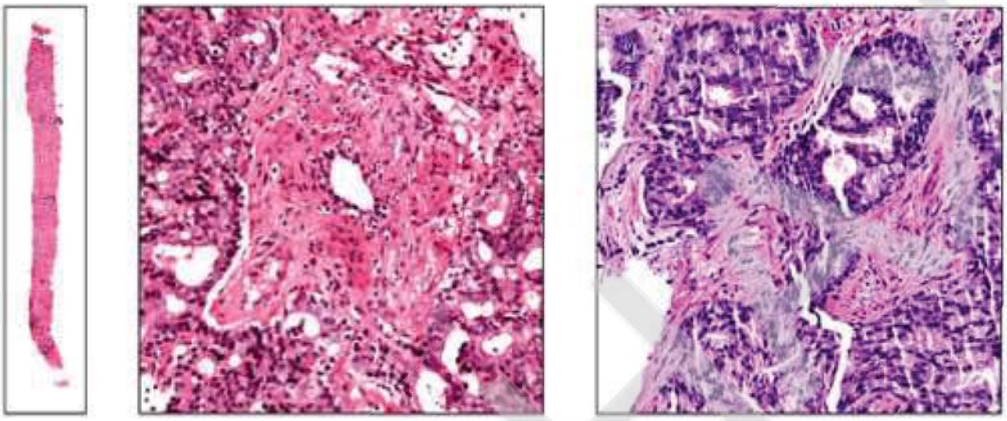
We developed two methods for the analysis of prostate WSI. The first one was based on the obtaining feature vectors for each tile of an image and then combining these vectors into single pseudo-image to be used for further prediction of ISUP score. This method is described in details in section 3.2. The second method works with canvases of tiles as a more compact representation of the whole WSI which can be used for training of deep learning models. Section 3.3 provide further explanation of this method.

3.1. Input Data

The original image dataset. The original dataset of histological images (acronym DS_RAW, see Table 1) consisted of 10 616 WSIs (see example in Fig. 1a) with the corresponding ISUP and Gleason scores. Images were provided by the Radboud University of Nijmegen, the Netherlands (source A, magnification x20, pixel spacing 0.24 μm, the second level with a pixel spacing of 0.48 μm was provided, see example in Fig.1b) and Karolinska Institute, Sweden (source B, magnification x20, pixel size 0.48 μm, Fig. 1c). Each WSI had a mask that indicated tumor areas and their Gleason

scores. We excluded 671 WSIs from the dataset since some artifacts were present or the mask and the scores did not match each other [8]. The cleanup resulted in the DS_CLEAN dataset. A test dataset included two private datasets of WSI. The first one had 420 samples in it and the second one consisted of 580 WSI that were inaccessible to anyone but organizers of the competition. Therefore, blind testing and assessment of developed methods were provided.

Inaccuracies in image labeling. In the process of labeling the input data, several labeling errors were made, due to which up to 30% of all WSIs had incorrect ISUP and Gleason grades [10].



a) *b)* *c)*

Fig. 1. Examples of original images:
a – WSI image (18 176 x 4 096 pixel size), b – tile from source A,
c – tile from source B

The main dataset of images (Table 1, acronym DS_MAIN). This dataset was used to train neural networks that determined a Gleason score of a single tile (see next section). To avoid errors in the Gleason score assignment due to the presence of two of Gleason score values for each WSI at the same time, we selected tiles only from WSIs, that had two equal grades, such as 3 + 3, 4 + 4, 5 + 5, as well as from the WSI containing healthy tissues, marked, accordingly, as 0 + 0. This approach revealed a lack of tiles from source B with a 5th Gleason score. To balance the dataset, we added Gleason 4 + 5 and 5 + 4 scored WSI tiles to 5th class. Consequently, the accuracy of separating the 4th and 5th classes decreased, but the value of the evaluation metric of the Kappa classes increased overall. Kappa can be defined as:

$$\kappa_w = 1 - \frac{q_0}{q_e},$$

$$q_0 = \sum_i \sum_j v_{ij} p_{ij},$$

$$q_e = \sum_i \sum_j v_{ij} p_i \, p_j \, ,$$

where p_i is a fraction of sample objects that the neural network categorized with the ISUP value of i, p_j is a fraction of objects of class j, p_{ij} is a fraction of objects belonging to category j that the neural network predicted as objects from category i, $v_{ij} = (i-j)^2$. We extracted all tiles from the WSI layer with magnification 4 times less than the maximum and with a size of 256x256 pixels. Information on the number of images in the datasets is presented in Tab. 1.

Table 1. Datasets description

Acronym	Number of WSIs	Number of tiles	Balance of ISUP scores, %						Balance of Gleason scores, %			
			0	1	2	3	4	5	0	3	4	5
DS_RAW	10616	-	27	25	13	12	12	11	18	51	45	13
DS_CLEAN	9945	-	27	25	13	12	12	11	18	51	45	13
DS_MAIN	5067	22089	51	32	-	-	15	2	43	25	19	13

Some tiles, as well as whole WSIs, can have areas with different Gleason scores, i.e. even 256x256 pixel images can simultaneously contain elements representing different classes.

As shown in Table 1, each WSI from DS_MAIN on average is represented by 4 tiles, which were not always contiguous. Overall, the width of the tissue sample obtained via needle biopsy of the prostate corresponds to only 1-3 tiles. Thus, extraction of overlapping tiles would not significantly increase the amount of data. Setting a lower threshold on the percentage of tissue that should be present on a tile might introduce additional inaccuracies due to the inevitable increase of the background on the tiles. The lack of overlapping tiles is partially compensated by their augmentation, which includes rotations and reflections in the process of training neural networks.

3.2. Sequential approach

The sequential approach to WSI analysis was based on the original WSI image being analyzed and classified tile by tile. Upon completion of the process, the obtained partial recognition results were generalized, so that a corresponding ISUP class could be assigned to the entire WSI image.

Firstly, we used ResNet50, MobileNetV2, VGG16 as baseline models for the tile classification problem. For training a model, we chose an RMSprop optimizer for all three networks as well as several data augmentation techniques such as reflection and rotation. Figure 1 shows a great example of difference among images from sources A and B, so to address this issue we applied colour normalization to all tiles using the Masenko method with masking [11]. Examples of normalized tiles from data sources A and B are shown in Fig. 2. The best results were achieved using the ResNet50 model, therefore, we used this architecture for all the subsequent experiments. The corresponding results are provided in Tab. 2.

Table 2. Results of preliminary experiments on tile classification

Network architecture, additional features	Accuracy	Weighted Kappa Metric
ResNet50	*0.87*	*0.87*
MobileNetV2	0.82	0.81
VGG16	0.68	0.71
ResNet50 (with MaxPooling)	0.84	0.83
ResNet50 (with a filter for misclassified tiles)	0.86	0.83

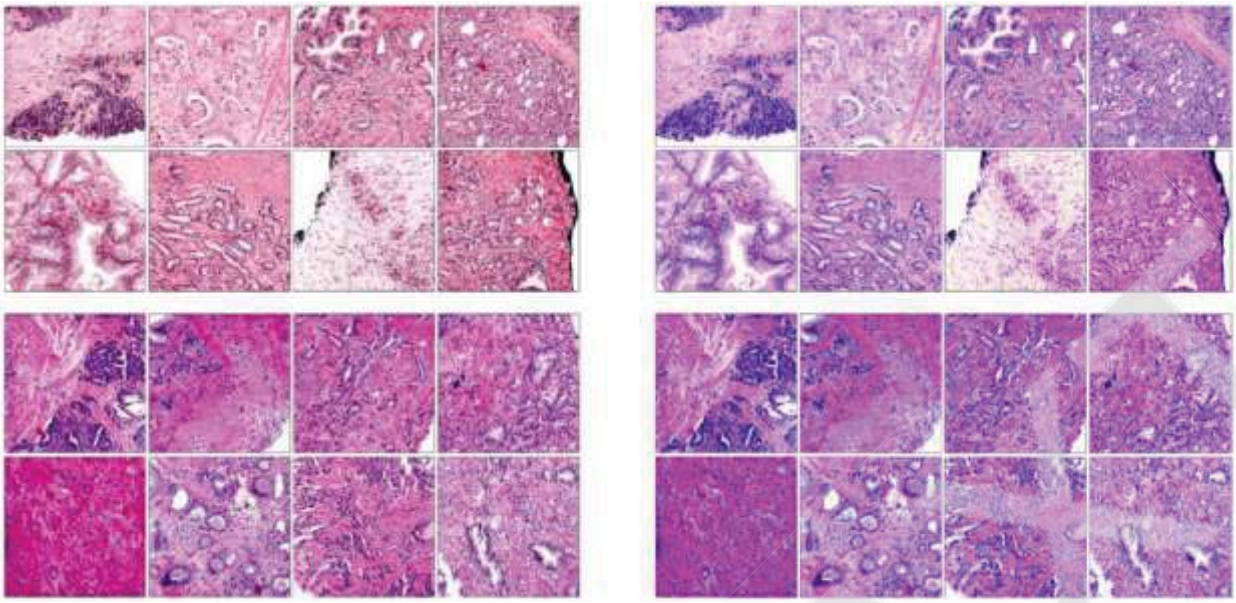


Fig. 2. Examples of original images (left) and their normalized versions (right). In both cases, the top 8 tiles come from source A and the bottom 8 from source B.

The attempt to train a network with kappa loss as a loss function did not allow us to obtain an accuracy higher than 0.59 due to high levels of noise in kappa loss. Due to the same reason and vast amounts of data combination of categorical cross-entropy and kappa losses did not show higher metric values than plain categorical cross-entropy. Because the trained model was not explicitly used for prediction but served as a feature extractor, we trained a model based on ResNet50 with a global pooling layer as penultimate to obtain one feature vector per tile. We tried both average and maximum poolings, however, average pooling showed better results in tile classification, as well as representations from this pooling, served better for ISUP score prediction than ones from maximum pooling.

Upon a detailed study of the test results, we discovered that there was a fairly large percentage of tiles among the test tiles obtained from source B, that were originally marked as cancerous, but were misclassified by the model as normal ones. The reasoning behind this could be the fact that only an approximate area of the tumour was provided by pathologists for images from source B, while the exact location of cancer cells was not known. Consequently, some of the selected from such areas tiles might not contain tissue samples with cancer. To minimize the effect of such labeling errors we used the data clearing technique. The essential part of this technique was predicting validation set with a trained neural network and then excluding tiles which were predicted as normal with high confidence when they were labeled as cancerous. This approach significantly increased recall for 0 Gleason score, but the value of this metric decreased for the rest of Gleason score classes.

After obtaining feature extractor for tiles, we used it to calculate feature vectors of size 2048 for each tile in WSI and fed them into the second generalizing neural network. The input for this generalizing CNN was a pseudo image which was created from tile feature vectors. In other words, each WSI corresponded to a pseudo-image of size WTxHTx2048, where WT and HT are the number of tiles in the width and height

of the WSI, respectively. The ultimate goal of CNN was to predict the ISUP score for WSI using tile features.

This approach was motivated by the fact that ISUP score is determined by the ratio of areas on the WSI with different Gleason scores in medical practice and the fact that spatial information must be essential for determining relationships among tiles

Since the spatial resolution of the resulting pseudo-images was too low, it was not possible to use all advantages and generalizing capabilities of such well-known neural networks as VGG16, ResNet, EfficientNet and others. Therefore, we propose a small network for this purpose, an architecture of which is described in detail in Table 3. ReLU function and batch normalization were used between all convolutional layers.

Table 3. Convolutional network used to predict ISUP metrics

Layer	**Number of input channels**	**Number of output channels**
Conv layer 3x3x2 + MaxPool	2048	1024
Conv layer 3x3x2 + MaxPool	1024	512
Conv layer 3x3x2 + MaxPool	512	256
Conv layer 3x3x2 + MaxPool	256	128
Dense layer	128	32
Dense layer	32	16
Dense layer	16	6

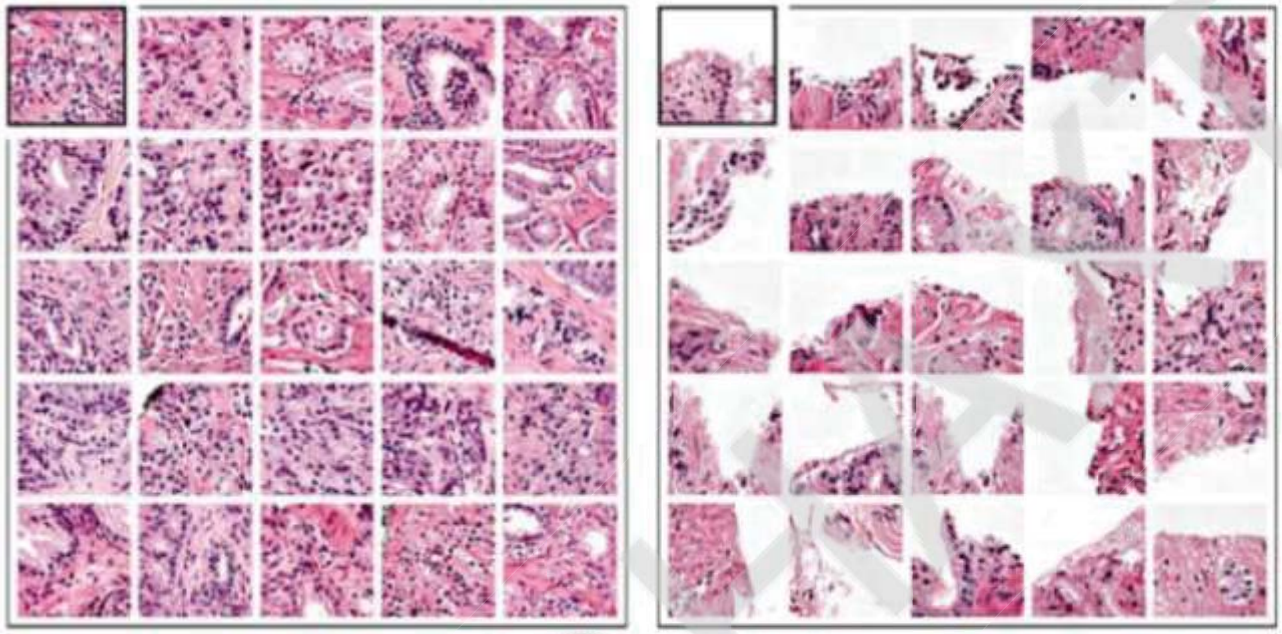
We performed several experiments on training the model with different loss functions, including categorical cross-entropy, kappa function as well as their combination. Due to the fact that categorical cross-entropy does not take into account the order in the distribution of classes, we did not achieve high metrics value on the test dataset. On the other hand, the Kappa function is rather noisy as a loss function, meaning the model might struggle to train properly. Therefore, it was decided to use a combined function, which is the sum of categorical cross-entropy and kappa function. In the process of training, the entropy quickly reaches its optimal value, after which the neural network optimizes mainly Kappa function. This fact allowed us to achieve the best accuracy for this method.

3.3. Canvas approach

Another way to preprocess WSI into data which can be used for CNN training and analysis was a conversion of a WSI to a canvas (mosaic) of tiles. Here a tile is defined as a small region of the image that contains a section of the tissue. The preprocessing includes the following three steps:

1. divide the WSI image into non-overlapping tiles so that each pixel of the image belongs to one tile only,
2. choose a predetermined number of some "representative" tiles among the obtained ones,
3. place representative tiles in a rectangle/square shaped collage and save it as a single image.

In order to form a canvas tiles containing the largest proportion (area) of prostate tissue were chosen as representative tiles. Examples of such tiles are shown in Fig. 3a, ones with a small area of a tissue are shown in Fig. 3d. The efficiency of this method can be explained by the specifics of the analyzed data: original WSI images had only a thin strip of prostate tissue, while most of the image was occupied by the background.



a b

Figure 3 Examples of tile canvases
a – tile with large area of tissue, b – tiles with small area of tissue

As a result of the described preprocessing method, each WSI image was associated with a much smaller image, which, nevertheless, had a sufficiently better representation of the diversity of tissue sites. The conversion from pairs "WSI - ISUP score" to pairs "canvas-tiles - ISUP score" allowed us to use existing deep learning methods for processing newly created canvases of tiles. Considering the presence of some ordering of the ISUP scale values, we addressed the problem of diagnosing prostate cancer as an ordinal regression, i.e. a label of each image was mapped into a vector of five numbers, so that M label was coded by setting the first M elements to 1 in the resulting vector, and the rest of the elements were equated to 0. For example, the value 0 was represented as [0, 0, 0, 0, 0] and the value 3 is like [1, 1, 1, 0, 0].

The resulting score was then predicted using deep learning techniques. We chose EfficientNet-B0 classifier due to its high efficiency, small number of parameters and relatively short training time. Since the problem was considered as an ordinal regression one, the proposed neural network model had 5 output values with sigmoid activation on last layer. Then obtained values were summed and rounded to obtain the final ISUP score. For example, a vector of predictions [0.9, 0.7, 0.8, 0.1, 0.2] corresponded to the value 3 of the ISUP scale. On the basis of the above, the binary cross-entropy was chosen as the loss function.

As indicated in table 4, changing the dimensions of the canvas image did not significantly affect the quality of the final results. In addition, we also tried to consider the problem as a classification problem, however, the results were significantly lower in this case.

Table 4. The results of the prediction of the ISUP scored by the image-canvas

	Input dataset	Tile canvas size	Weighted Kappa Metric
1	DS_CLEAN	4x4	0.884
2	DS_CLEAN	5x5	0.892
3	DS_CLEAN	6x6	0.881
4	DS_RAW	5x5	0.883
5	DS_RAW	6x6	0.883

To further improve out solution we used several techniques which are listed below.

(a) Building an ensemble of models. A certain subset of trained neural networks was selected. Each model from the subset predicted ISUP score of each test image, so that the final answer was a mean of their predictions. It should be mentioned that there are more complex methods for building an ensemble of models, for example, using a weighted average, model stacking, etc. However, we found the chosen method to be efficient enough for the studied problem.

(b) Shifted tile slicing. In addition to slicing tiles from the upper left corner of the WSI, slicing was also done by shifting a start point by 64/128/192 pixels down and to the left. After each shift, we used a new canvas image to perform prediction. This resulted in 4 different predictions that were avagered to get a final score. Described approach allows to find various image texture features might be missed due to a single fixed cut.

(c) Augmentation of a test dataset. Recently, it has become a very widespread method of stabilizing predictions. The idea is to transform images and predict scores both for the initial test dataset and the transformed one with subsequent averaging of the results. The purpose of this method is to reduce the likelihood of overfitting by capturing any visual image features that are not important.

As a final solution we build an ensemble of models listed as 1, 3, 4 and 5 in Table 4. The second model did not improve the results obtained by an ensemble, therefore, it was not used for the final ensemble. Augmentation was performed in a manner similar to one described earlier by reflecting and rotating images.

IV. RESULTS

The developed method with sequential classification of WSI tiles for further use of extracted features for ISUP prediction achieved kappa-metric of 0.807 on the available test sample and 0.811 on the closed to public test sample. Another method with straightforward prediction of ISUP score based on tile canvases as input for ensemble of 4 convolutional neural networks achieved 0.8863 on a public test dataset and 0.9277 on a private test dataset. In addition to this ensemble of two neural networks (Table 4, networks 1 and 3) achieved similar figures for kappa metric of 0.8836 and

0.9297 on public and private test datasets respectively. For better comparison they are listed in the Table 5.

Table 5. Results of predicting ISUP score for WSI from public and private test sets.

Quadratic kappa metric	Public test set	Private test set
Sequential approach	0.807	0.811
Canvas approach (ensemble 1, 3, 4, 5)	0.886	0.928
Canvas approach (ensemble 1, 3)	0.884	0.930

V. CONCLUSIONS

A comparative analysis of the methods studied in this work allows us to draw the following conclusions:

(1) The approach of sequential recognition of image fragments (tiles) with subsequent generalization of the results performed relatively poorly compared to the alternative approach - recognition of canvases of representative tiles.

(2) The presence of certain order relations of the output classes such as disease stages should be taken into account when deciding whether a problem should be viewed as a regression or classification one.

(3) Accounting the specifics of the input data and replacing whole-slide images with a subset of representative tiles made it possible to build efficient neural network models despite a significant loss of information.

(4) The techniques of prediction stabilization, such as ensembles of neural networks and augmentation of test data, used in this work, resulted in increase of the prediction accuracy, while training more complex neural network models mostly led to overfitting.

VI. REFERENCES

1. Rawla, P. (2019). Epidemiology of prostate cancer. *World journal of oncology*, *10*(2), 63.

2. Epstein, J. I., Allsbrook Jr, W. C., Amin, M. B., Egevad, L. L., & ISUP Grading Committee. (2005). The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, *29*(9), 1228-1242.

3. Camparo, P., Egevad, L., Algaba, F., Berney, D. M., Boccon- Gibod, L., Compérat, E., ... & Varma, M. (2012). Utility of whole slide imaging and virtual microscopy in prostate pathology. *Apmis*, *120*(4), 298-304.

4. Goldenberg, S. L., Nir, G., & Salcudean, S. E. (2019). A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology*, *16*(7), 391-403.

5. Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M., ... & Eklund, M. (2020). Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, *21*(2), 222-232.

6. Pantanowitz, L., Quiroga-Garza, G. M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C., ... & Dhir, R. (2020). An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health*, *2*(8), e407-e416.

7. Nagpal, K., Foote, D., Liu, Y., Chen, P. H. C., Wulczyn, E., Tan, F., ... & Stumpe, M. C. (2019). Development and validation of a deep learning algorithm for improving Gleason scoring of

prostate cancer. *NPJ digital medicine*, *2*(1), 1-10.

8. Bulten, W., Balkenhol, M., Belinga, J. J. A., Brilhante, A., Çakır, A., Egevad, L., ... & Litjens, G. (2020). Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Modern Pathology*, 1-12.

9. Luan, S., Chen, C., Zhang, B., Han, J., & Liu, J. (2018). Gabor convolutional networks. *IEEE Transactions on Image Processing*, *27*(9), 4357-4366.

10. Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*(2), 243-253.

11. Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., ... & Thomas, N. E. (2009, June). A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (pp. 1107-1110). IEEE.