Intelligent information systems

UDC 004.85 **doi: 10.20998/2522-9052.2019.4.10**

D. Hlavcheva, V. Yaloveha, A. Podorozhniak

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine

APPLICATION OF CONVOLUTIONAL NEURAL NETWORK FOR HISTOPATHOLOGICAL ANALYSIS

Abstract. Among all types of cancer, breast cancer is the most common. In 2017 breast cancer was the fourth rate for death reasons in Ukraine. The paper is devoted to the automatization of histopathological analysis, which can improve the process of cancer stage diagnosis. **The purpose** of the paper is to research the ability to use convolutional neural networks for classifying biopsy images for cancer diagnosis. **The tasks** of research are: analyzing cancer statistics in Europe and Ukraine; analyzing usage of Machine Learning in cancer prognosis and diagnosis tasks; preprocessing of BreCaHAD dataset images; developing a convolutional neural network and analyzing results; the building of heatmap. **The object** of the research is the process of detecting tumors in microscopic biopsy images using Convolutional Neural Network. **The subject** of the research is the process of classifying healthy and cancerous cells using deep learning neural networks. **The scientific novelty** of the research is using ConvNet trained on the BreCaHAD dataset for histopathological analysis. The theory of deep learning neural networks and mathematical statistics methods are used. **In result** it is obtained that the classification accuracy for a convolutional neural network on the test data is 0.935, ConvNet was effectively used for heatmap building.

Keywords: deep learning; convolutional neural networks; breast cancer; histopathological analysis; biopsy images; BreCaHAD.

Introduction

Among all types of cancer, breast cancer is the most common. Breast cancer is the second-highest mortality rate after lung and bronchial cancer, and about 30% of newly diagnosed cases are breast cancer [1].

Europe contains 9% of the world population but has a 25% share of the global cancer burden [2]. According to [3] in 2018 the number of new cancer cases is more than 4 million for both sexes and all ages. The first rank of incidence is referring to breast cancer (more than 520 thousand). Total mortality from cancer in 2018 in Europe is almost 2 million for both sexes and all ages. The first rank of mortality is referring to lung cancer (almost 390 thousand). The number of prevalent cases for the 5-year period is more than 12 million for both sexes and all ages.

In Ukraine [4] the number of new cases in 2018 is almost 170 thousand for both sexes and all ages. The number of death is almost 100 thousand. The top types of cancer by incidence and mortality in Ukraine is consistent with Europe.

According to [5] breast cancer is the fourth rate in death reasons in Ukraine. In 2017 the age-adjusted Death Rate is 20.79 per 100,000 of population ranks Ukraine is rank #48 in the world.

The big amount of cases causes the appearance of large amounts of data, the analysis of which can be used for diagnosis, understanding the causes of diseases, patients risk groups, etc.

Nowadays, Machine Learning is widely used for big data structures processing. The ability of ML tools to detect key features in complex data sets shows their importance. Various methods, including artificial neural networks (ANN), Bayesian networks (BN), Support vector machines (SVMs) and Decision Trees (DTs) are widely used in cancer research and helps in efficient and accurate decision making [6].

In [7] breast cancer survival predicted by the DT algorithm. It was trained on the SEER database [8] (200 thousand cases have been used) and resulted accuracy was 0.93. In [9] Multiple myeloma cancer susceptibility predictions have been made on SNPs data [10] with 0.71 resulted accuracy.

ML technics are being used mostly for prediction tasks. Processing of clinical images (results of different diagnostics types) needs ANN, particularly, Deep Learning.

Convolutional Neural Network (ConvNet) for brain tumor detection used in [11]. It processes MRI images and gets 0.99 train accuracy and 0.986 validation accuracy. In [12] Computer Tomography images have been processed by 3D - ConvNet and result AUC was 0.83.

For breast cancer diagnosis biopsy images, thermal images and mammography images analyzing are being used. In [13] Multi-Scale CNN has been used for mammography classification (Digital Database for Screening Mammography [14]) and get AUC = 0.92. The study [15] presents a computer-aided diagnosis system based on convolutional neural networks as an alternative diagnosis methodology for breast cancer diagnosis with thermal images. Resulted accuracy for CNN was 0.86.

The purpose of the paper is to research the ability to use convolutional neural networks for classifying biopsy images for cancer diagnosis.

In the paper, the Convolutional Neural Network is used to solve the problem of detecting healthy and cancerous cells in microscopic biopsy images. The heatmap will be built. Such an approach can improve the process of cancer stage diagnosis.

1. Convolutional Neural Networks

Convolutional Neural Network it is a deep learning approach for data, particulary image, processing.

Deep Learning concept was first known as hierarchical learning at the [16].

Deep learning uses data processing by several layers of neurons in such a way that each layer highlights certain features and gives it to the next layer for processing. Such an architecture makes it possible to generalize input data [17]. Until 2012 in the field of computer vision, neural networks were better known for their tendency for overfitting than the ability to solve complex problems of visual recognition. In 2012 ConvNets show good result in ImageNet competition [18] and attracted attention to their usage for image processing [19].

Convolutional neural network was developed by Yann LeCun in 1988 [20]. ConvNet were created on the basis of “simple cells” in human brain. Such cells were discovered in 1960s by Torsten Nils Wiesel and David Habel [21].

The architecture of a typical ConvNet has several parts. The first few parts consist of two types layers: convolutional layers and pooling layers. Convolutional layers consist of feature cards, each of them has a set of weights called a filter bank. The local weighted sum at the output of these 2 layers is then transmitted via non-linearity, such as ReLu. Each feature map with its own filter bank by the sliding window method is applied to the entire input image.

Firstly, such an architecture makes it possible to identify local features. Secondly, the removal of features does not depend on the location of the objects. In other words, an object can appear in any part of the image and, thanks to feature maps, will be found. The mathematical operation performed is a discrete convolution [22].

Convolution is a mathematic operation on two functions $f(x)$ and $g(x)$ that produces a third function. Having two-dimensional image, I and array K with size h×w (also called as convolution kernel) convolute image $I*K$ is calculated by putting the kernel on image. The sum of the multiplied elements of output image and kernel is written [23]:

$$(I * K)_{xy} = \sum_{i=1}^{h} \sum_{j=1}^{w} K_{ij} \times I_{x+i-1,y+j-1} \,. \quad (1)$$

Although the role of the convolutional layer is to detect local conjunctions of features from the previous layer, the role of the pooling layer is to merge semantically similar features into one. The most used pooling type is MaxPooling [22].

Pairs of Convolutional and Pooling layers followed by several Fully Connected layers.

ConvNet has tendency to overfitting. There are some methods to avoid it.

Dropout prevents overfitting and provides a way of approximately combining exponentially many different neural network architectures efficiently. The term “dropout” refers to dropping out units (hidden and visible) in a neural network [24]. In this layer with probability of p the neuron is excluded from the network at the time of current iteration [25].

Data augmentation is a key element in training high-dimensional models. In this approach, one synthesizes new observations by applying pre-specified transformations to the original training data. In practice, data augmentation is a manual process, where a human specifies a small set of transformations; for image classification tasks, these are most commonly chosen to be simple linear transformations such as translations, rotations and scaling. [26].

2. Data

Histopathological tissue analysis by a pathologist plays an important role in the diagnosis and prognosis of breast cancer. The diagnosis of biopsy tissue with hematoxylin and eosin-stained images is non-trivial and specialists often disagree on the final diagnosis [27].

In the research, BreCaHAD (breast cancer histopathological annotation and diagnosis dataset) has been used [28]. This dataset provides 162 breast cancer histopathology images. The dataset includes various malignant cases. The task for researchers, proposed for this dataset is automated classifying histological structures in hematoxylin and eosin (H&E) stained images into six classes: tumor nuclei, non-tumor nuclei, apoptosis, mitosis, tubule, and non-tubule.

The BreCaHAD dataset contains microscopic biopsy images in uncompressed (.TIFF) image format, three-channel RGB with 8-bit depth in each channel, and the dimension is 1360 × 1024 pixel and each image is annotated. Anotation corresponds to the marked objects of 6 classes and provided in json format [29].

Example of image with annotated objects is presented in Fig. 1.

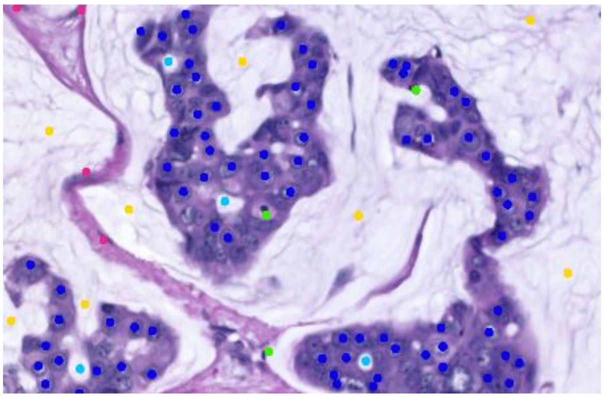

Fig. 1. Annotated image from BreCaHAD dataset

For Convolutional Neural Network images from BreCaHAD have been sliced into 32×32 pixel size. All images were divided into 2 groups:

- Tumor: according to “tumor” objects in annotations;
- Non-tumor: image fragments where annotated objects aren’t appearing.

Total number of images was 40000. Sample images for both groups are shown in Fig. 2.

All images were divided into 3 datasets:

- Training – 14000 images (70%);
- Validation – 3000 images (15%);
- Test – 3000 images (15%).

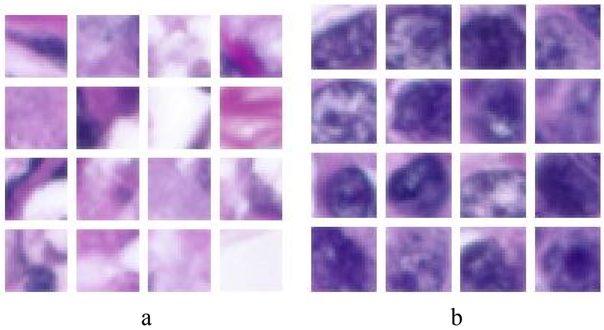

Fig. 2. Examples of ConvNet input images
(a – tumor, b – non-tumor)

Each image has been normalized: each pixel value was correlated to the gap [0;1].

3. Results and Discussions

For developing ConvNet Python 3.6 was used as programming language. For Neural Network compiling, training and evaluating Keras version 2.2.4 with Tensorflow backend version 1.12.0 was used.

For experiment Convolutional Neural Network has been compiled. It's structure is shown at the Table 1. Total number of parameters – 171105. ReLu activation function was used in ConvNet. As optimization algorithm Adam was used.

Table 1 – Structure of the experiment

Type	Input	Output	Para-meters
1. Convolutional	(32x32x3)	(30x30x32)	896
2. MaxPooling	(30x30x32)	(15x15x32)	0
3. Convolutional	(15x15x32)	(14x14x32)	4128
4. MaxPooling	(14x14x32)	(14x14x32)	0
5. Convolutional	(12x12x64)	(12x12x64)	18496
6. MaxPooling	(12x12x64)	(6x6x64)	0
7. Flattening	(6x6x64)	2304	0
8. Fully Connected	2304	64	16448
9. Dropout	64	64	0
10. Fully Connected	64	1	65

ConvNet was trained during 30 epochs with batch size - 16. Accuracy and loss on training data is shown in Fig. 3, on validation data – in Fig. 4.

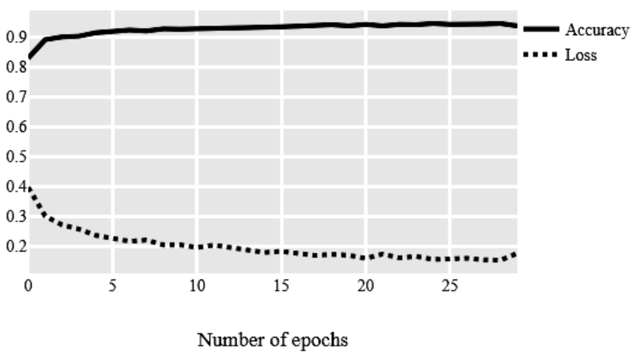

Fig. 3. Accuracy and loss on training data

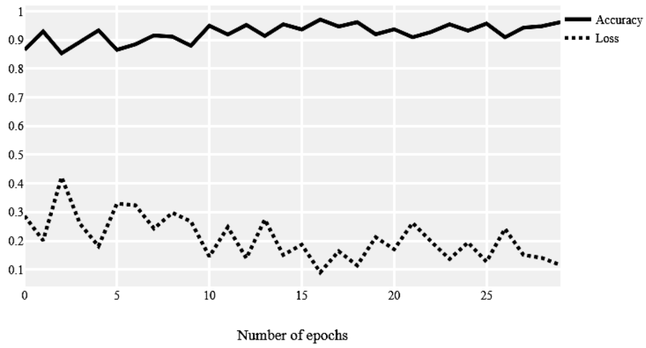

Fig. 4. Accuracy and loss on validation data

Accuracy on train data is 0.935. It is obtained that ConvNet can effectively classify objects in this task.

Heatmap for full image from dataset (1036×1024 pixel) was built. Results are shown on Fig. 5. Heatmap was built using floating window method with step size 8 pixels.

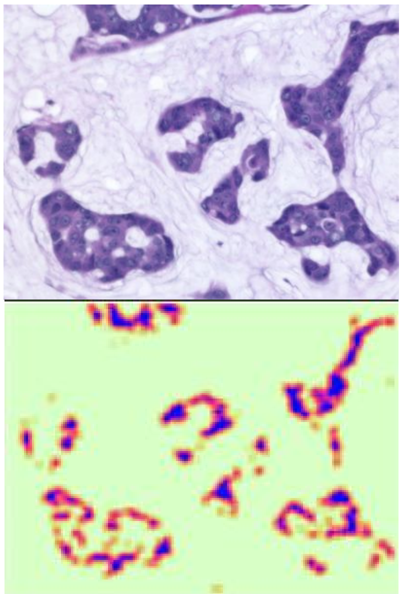

Fig. 5. Heatmap

In [27] authors used 512×512 pixel with 3 RGB-channels biopsy images as input data for ConvNet. ConvNet has 4 pairs of convolutional and maxpooling layers and 3 fully-connected layer. And get result 0.778 for four classes classification. In our work accuracy is higher because of less number of classes. Binary classification was held in [1]. CNN classified biopsy images 350×230 to Benign or Malignant with accuracy 0.934. ConvNet has almost the same architecture as in our research, but with bigger amount of feature maps in convolutional layers. Similarity in architecture and the same task (binary classification) is the reason of similar results.

Conclusions

In the world breast cancer is the most common. In Ukraine, in 2018 the number of new breast cancer cases is almost 19 thousand, and the number of death is more than 8 thousand [4].

The Machine Learning approach is widely used in cancer prediction and diagnosis researches. In the paper examples of cancer prediction and medical images analysis are shown. ConvNet is an effective instrument for image classification, but it is important to pay attention to the overfitting problem. BreCaHAD dataset can be effectively used for solving the task of histopathological analysis automatization.

In this paper, Convolutional Neural Networks was used for microscopic biopsy images classifying. Classification accuracy on test data obtained 0.935. We showed that trained ConvNet can be used for heatmap building. Such an approach can improve the process of cancer stage diagnosis. Results have been compared with related works.

In future work, we are going to increase dimensions of input images, make multiclass classification, pay more attention to overfitting problem and use a detection algorithm that determines the specific location of objects with specific probability.

References

1. Dabeer, S., Khan, M. M. and Islam, S. (2019), "Cancer diagnosis in histopathological image: CNN based approach", *Informatics in Medicine Unlocked*, 16, article 100231.
2. Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., Gavin, A., Visser, O. and Bray, F. (2018), "Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018", *European Journal of Cancer,* 103, pp. 356-387.
3. *International Agency for Research of Cancer, World Health Organization Europe: Europe – Global Cancer Observatory* (2018), available at: https://gco.iarc.fr/today/data/factsheets/populations/908-europe-fact-sheets.pdf.
4. *International Agency for Research of Cancer, World Health Organization Europe: Ukraine – Global Cancer Observatory* (2018), available at: https://gco.iarc.fr/today/data/factsheets/populations/804-ukraine-fact-sheets.pdf.
5. *World health rankings: Breast Cancer in Ukraine* (2017), available at: https://www.worldlifeexpectancy.com/ukraine-breast-cancer.
6. Konstantina Kourou, Themis P.Exarchos, Konstantinos P.Exarchos, Michalis V.Karamouzis and Dimitrios I.Fotiadi (2014), "Machine learning applications in cancer prognosis and prediction", *Computational and structural biotechnology journal,* 13, pp. 8-17.
7. Delen, D., Walker, G. and Kadam, A. (2005), "Predicting breast cancer survivability: a comparison of three data mining methods", *Artificial intelligence in medicine,* 34(2), pp. 113-127.
8. *SEER Incidence Database - SEER Data & Software* (2019), available at: https://seer.cancer.gov/data/.
9. Waddell, M., Page, D. and Shaughnessy Jr, J. (2018), "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma", *Proceedings of the 5th international workshop on Bioinformatics*: ACM, pp. 21-28.
10. *SNP – NCBI* (2019), available at: https://www.ncbi.nlm.nih.gov/snp/.
11. Sawant, A., Bhandari, M., Yadav, R., Yele, R. and Bendale, M.S. (2018), "Brain cancer detection from mri: A machine learning approach (tensorflow)", *Brain,* 5(04).
12. Alakwaa, W., Nassef, M. and Badr, A. (2017), "Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)", *Lung Cancer, 8*(8), pp. 409.
13. Lotter, W., Sorensen, G. and Cox, D. (2017), "A multi-scale CNN and curriculum learning strategy for mammogram classification", *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*: Springer, pp. 169-177.
14. Heath, M., Bowyer, K., Kopans, D., Moore, R. and Kegelmeyer, W. P. (2000), "The digital database for screening mammography", *Proceedings of the 5th international workshop on digital mammography*: Medical Physics Publishing, pp. 212-218.
15. Zuluaga-Gomez, J., Masry, Z. A., Benaggoune, K., Meraghni, S. and Zerhouni, N. (2019), "A CNN-based methodology for breast cancer diagnosis using thermal images", *arXiv preprint arXiv:1910.13757.*
16. Bengio, Y. (2009), "Learning deep architectures for AI", *Foundations and trends® in Machine Learning,* 2(1), pp. 1-127.
17. Kuchuk, H., Podorozhniak, A., Hlavcheva, D. and Yaloveha, V. (2019), "Application of Deep Learning in the Processing of the Aerospace System's Multispectral Images", *Handbook of Research on Artificial Intelligence Applications in the Aviation and Aerospace Industries*, IGI Global, pp. 134-147.
18. *Imagenet large scale visual recognition challenge* (2013), available at: http://www.image-net.org/challenges/LSVRC/2013/.
19. Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A. and Carlsson, S. (2015), "'Factors of transferability for a generic convnet representation'", *IEEE transactions on pattern analysis and machine intelligence,* 38(9), pp. 1790-1802.
20. LeCun, Y. (1989), "Generalization and network design strategies", *Connectionism in perspective*: Citeseer.
21. Hubel, D. H. and Wiesel, T. N. (1962), "'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex'", *The Journal of physiology,* 160(1), pp. 106-154.
22. LeCun, Y., Bengio, Y. and Hinton, G. (2015), "Deep learning", *nature*, No. 521 (7553), pp. 436-444, available at: https://doi.org/10.1038/nature14539.
23. Liubchenko, N., Podorozhniak, A. and Bondarchuk, V. (2017), "Neural network method of intellectual processing of multispectral images", *Advanced Information Systems*, Vol. 1, No. 2, pp. 39-44, DOI: https://doi.org/10.20998/2522-9052.2017.2.07.
24. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014), "Dropout: a simple way to prevent neural networks from overfitting", *The journal of machine learning research,* 15(1), pp. 1929-1958.
25. Yaloveha, V., Hlavcheva, D., Podorozhniak, A. and Kuchuk, H. (2019), "Fire Hazard Research of Forest Areas based on the use of Convolutional and Capsule Neural Networks", *2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, IEEE, pp. 828-832.
26. Hauberg, S., Freifeld, O., Larsen, A. B. L., Fisher, J. and Hansen, L. (2010), "Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation", *Artificial Intelligence and Statistics*, pp. 342-350.
27. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A. and Campilho, A. (2017), "Classification of breast cancer histology images using convolutional neural networks", *PloS one,* 12(6), article e0177544.

28. *BreCaHAD: A Dataset for Breast Cancer Histopathological Annotation and Diagnosis* (2019), available at: https://figshare.com/articles/BreCaHAD_A_Dataset_for_Breast_Cancer_Histopathological_Annotation_and_Diagnosis/7379186.
29. Aksac, A., Demetrick, D. J., Ozyer, T. and Alhajj, R. (2019), "BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis", *BMC research notes,* 12(1), pp. 82.

Надійшла (received) 29.09.2019
Прийнята до друку (accepted for publication) 20.11.2019

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Главчева Дар`я Максимівна – студентка кафедри обчислювальної техніки та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Daria Hlavcheva – student of Computer Science and Programming Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: agafina99@gmail.com; ORCID ID: http://orcid.org/0000-0001-6990-6845

Яловега Владислав Анатолійович – магістрант кафедри обчислювальної техніки та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Vladyslav Yaloveha – master student of Computer Science and Programming Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: vladyslavyaloveha@gmail.com; ORCID ID: http://orcid.org/0000-0001-7109-9405

Подорожняк Андрій Олексійович – кандидат педагогічних наук, доцент, доцент кафедри обчислювальної техніки та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Andrii Podorozhniak – Candidate of Pedagogical Sciences, Associate Professor, Associate Professor of Computer Science and Programming Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: andriipodorozhniak@gmail.com; ORCID ID: http://orcid.org/0000-0002-6688-8407

Використання згорткових нейронних мереж для гістопатологічного аналізу

Д. М. Главчева, В. А. Яловега, А. О. Подорожняк

Анотація. Серед усіх видів раку найпоширенішим є рак молочної залози. У 2017 році рак молочної залози став четвертою причиною смертності в Україні. Стаття присвячена автоматизації гістопатологічного аналізу, що може покращити процес діагностики стадії раку. **Мета статті** - дослідити можливість використання згорткових нейронних мереж для класифікації зображень біопсії для діагностики раку. Відповідно до мети поставлено такі **завдання**: аналіз статистики захворюваності на рак в Європі та Україні; аналіз використання машинного навчання для завдань прогнозування та діагностики раку; попередня обробка зображень набору даних BreCaHAD; навчання згорткової нейронної мережі та аналіз результатів; побудова теплової карти. **Об'єктом** дослідження є процес виявлення пухлин на мікроскопічних зображеннях біопсії за допомогою згорткових нейронних мереж. **Предметом** дослідження є процес класифікації здорових та ракових клітин за допомогою нейронних мереж глибокого навчання. **Науковою новизною** дослідження є використання згорткової нейронної мережі, навченої на наборі даних BreCaHAD для виконання гістопатологічного аналізу. Використовуються теорія нейронних мереж глибокого навчання та методи математичної статистики. **В результаті** отримано, що точність класифікації згорткової нейронної мережі за тестовими даними становить 0,935, ця мережа може бути ефективно використана для побудови теплової карти.

Ключові слова: глибоке навчання; згорткові нейронні мережі; рак молочної залози; гістопатологічний аналіз; біопсія; BreCaHAD.

Использование сверточных нейронных сетей для гистопатологического анализа

Д. М. Главчева, В. А. Яловега, А. О. Подорожняк

Аннотация. Среди всех видов рака наиболее распространенным является рак молочной железы. В 2017 году рак молочной железы стал четвертой причиной смертности в Украине. Статья посвящена автоматизации гистопатологического анализа, что может улучшить процесс диагностики стадии рака. **Цель** статьи - исследовать возможность использования сверточных нейронных сетей для классификации изображений биопсии для диагностики рака. В соответствии с целью поставлены следующие **задачи**: анализ статистики заболеваемости раком в Европе и Украине; анализ использования машинного обучения для задач прогнозирования и диагностики рака; предварительная обработка изображений набора данных BreCaHAD; обучение сверточной нейронной сети и анализ результатов; построение тепловой карты. **Объектом** исследования является процесс выявления опухолей на микроскопических изображениях биопсии с помощью згортковых нейронных сетей. **Предметом** исследования является процесс классификации здоровых и раковых клеток с помощью нейронных сетей глубокого обучения. **Научной новизной** исследования является использование сверточной нейронной сети, обученной на наборе данных BreCaHAD для выполнения гистопатологического анализа. Используются теория нейронных сетей глубокого обучения и методы математической статистики. **В результате** получено, что точность классификацией згортковых нейронной сети с тестовыми данными составляет 0,935, эта сеть может быть эффективно использована для построения тепловой карты.

Ключевые слова: глубокое обучение; сверточные нейронные сети; рак молочной железы; гистопатологический анализ; биопсия; BreCaHAD.