UDC 004.932.2:616-006.6:681.5.015
DOI https://doi.org/10.32782/tnv-tech.2025.1.12

BREAST CANCER DETECTION IN HISTOPATHOLOGY IMAGES USING SWIN V2 AND THE INFORMATION EXTREME METHOD

Papchenko O. *– Postgraduate Student at the Department of Computer Science, Sumy State University*
ORCID ID: 0009-0005-9918-566X
Researcher ID: JMC-1046-2023

Kuzikov B. O. *– PhD in Technical Sciences, Associate Professor at the Department of Computer Science, Sumy State University*
ORCID ID: 0000-0002-9511-5665
Scopus-Author ID: 55653809800
Researcher ID: AAS-7954-2020

The article addresses the application of machine vision in the analysis of histopathological images. The objective of the study is to improve the accuracy of automatic breast cancer detection in histopathological images by developing and implementing novel hybrid architecture that combines modern visual transformers with the information-extreme intellectual technology.

The paper presents a comparative analysis of the effectiveness of different neural network architectures for solving the problem of binary classification of histopathological images. Two fundamentally different approaches were investigated: Convolutional Neural Networks (CNN) based on ResNet architecture and Visual Transformers (ViT) based on the Swin Transformer V2 architecture. These approaches were used as base models in combination with the Information Extreme Technology (IET) for image classification.

The focus is on the model based on Swin Transformer V2 (SwinV2). SwinV2 employs an innovative attention mechanism in fixed windows with shifting, which ensures linear computational complexity relative to the image size, as opposed to quadratic complexity with global attention. The developed model utilizes a large SwinV2 architecture with 3 billion parameters, pre-trained on the extensive ImageNet-22K dataset, followed by fine-tuning on the specialized BreakHis dataset.

Experimental studies conducted on a balanced test set (847 samples for each class) demonstrate that the proposed approach using SwinV2 and IET achieves a classification accuracy of 98.5%, which is 10% higher than the results of a similar system based on ResNet (88.98%). This significant improvement is attributed to the ability of transformers to more effectively process global dependencies and object shapes in images, which is particularly important when analyzing the morphology of cell nuclei in histopathological images.

The study analyzed the key differences between CNN and ViT architectures, specifically their different biases towards textures and shapes when processing images. It was established that CNNs exhibit a strong inclination towards analyzing local texture patterns, while ViTs perform significantly better with global information about object shapes.

Based on the research results, promising directions for further studies were outlined, including the development of ensemble methods that combine the advantages of both architectures to create more reliable diagnostic systems with high accuracy. The proposed approach can be adapted for the analysis of other types of histopathological images and integrated into existing computer-aided diagnostic systems.

Key words: *histopathological images, convolutional neural networks, visual transformers, computer-aided diagnosis, medical image analysis, image classification, information-extreme technology.*

Папченко О., Кузіков Б. О. Виявлення раку молочної залози на гістопатологічних зображеннях з використанням Swin v2 та інформаційно-екстремальної технології

Стаття розглядає питання застосування машинного зору в задачах аналізу гістопатологічних знімків. ***Мета*** *дослідження полягає у підвищенні точності автоматичного детектування раку молочної залози на гістопатологічних зображеннях шляхом розробки*

та впровадження нової гібридної архітектури, що поєднує сучасні візуальні трансформери із інформаційно-екстремальною інтелектуальною технологією.

У роботі представлено порівняльний аналіз ефективності різних архітектур нейронних мереж для вирішення задачі бінарної класифікації гістопатологічних зображень. Досліджено два фундаментально різних підходи: згорткові нейронні мережі (Convolutional Neural Networks, CNN) на базі архітектури ResNet та візуальні трансформери (Visual Transformers, ViT) на базі архітектури Swin Transformer V2. Ці підходи використані як базові моделі у поєднанні з інформаційно-екстремальною технологією (Information Extreme Technology, IET) для класифікації зображень.

Основну увагу приділено моделі на базі Swin Transformer V2 (SwinV2). SwinV2 застосовує інноваційний механізм уваги у фіксованих вікнах із їх зміщенням, що забезпечує лінійну складність обчислень відносно розміру зображення, на відміну від квадратичної складності при глобальній увазі. Розроблена модель використовує попереднє навчання на масштабному наборі даних ImageNet-22K з подальшим точним налаштуванням на спеціалізованому наборі даних BreakHis.

Проведені експериментальні дослідження на збалансованому тестовому наборі (847 зразків для кожного класу) показують, що запропонований підхід з використанням SwinV2 та IET досягає точності класифікації 98.5%, що на 10% перевищує результати аналогічної системи на основі ResNet (88.98%). Це суттєве покращення пояснюється здатністю трансформерів ефективніше обробляти глобальні залежності та форми об'єктів на зображеннях, що особливо важливо при аналізі морфології ядер клітин у гістопатологічних зображеннях.

У процесі дослідження проаналізовано ключові відмінності між CNN та ViT архітектурами, зокрема їх різні упередження щодо текстур та форм при обробці зображень. Встановлено, що CNN демонструють сильну схильність до аналізу локальних текстурних патернів, тоді як ViT значно краще працюють з глобальною інформацією про форму об'єктів.

За результатами дослідження окреслено перспективні напрямки подальших досліджень, зокрема розробку ансамблевих методів, що поєднують переваги обох архітектур для створення більш надійних систем діагностики з високою точністю. Запропонований підхід може бути адаптований для аналізу інших типів гістопатологічних зображень та інтегрований у існуючі системи комп'ютерної діагностики.

Ключові слова: *гістопатологічні зображення, згорткові нейронні мережі, візуальні трансформери, комп'ютерна діагностика, аналіз медичних зображень, класифікація зображень, інформаційно-екстремальна технологія.*

Introduction. Histological image-based cancer diagnosis continues to be the standard for diagnosing cancer [1]. However, current computer technology for this task lags behind clinical needs [2]. Manual analysis of histological tissues is still the main diagnostic method and is largely dependent on the skills and experience of histopathologists. Such manual intervention has disadvantages compared to CAD systems [3]:

- It is time-consuming and difficult to evaluate in a reproducible way.
- The manual inspection is error prone due to the difficulty of the task.
- It is empirically known that there is significant variation between experts both within and between observations.

With the gradual development of artificial intelligence for CAD systems, many machine learning methods have been applied. This technology has the potential to exceed human performance over time and learn more efficiently. In such a way, integrating machine learning into the diagnostic procedure can have a positive effect, helping pathologists evaluate and analyze huge amounts of medical data [4]. It can also speed up the process by being able to process large data sets much faster [5]. Taking into account the importance of the task, the goal of this work is to compare functional efficiency of convolutional based neural networks (CNN) and visual transformer based (ViT) approaches used as backbones with Information Extreme Technology (IET) used as classifier applied in tasks of breast cancer detection on histopathological images.

Analysis of subject area. CAD systems for histological images analysis have a long history of improvement and development [6]. Modern computer vision systems are based on state of art approaches such as CNN or ViT [7].

While analyzing approaches of histopathology imaging cancer diagnosis systems, it is important to realize the main challenges which arise in this particular field. Factors that hinder the development of effective computer-aided histological analysis include:

- the great diversity and high complexity of histological features make it difficult to develop a universal computer system for analyzing images of different types of cancer;
- widespread image processing systems for radiology and cytology cannot be directly used for histological images due to different imaging technologies and different characteristics of these image types, transfer learning in this case gives mixed results;
- the lack of high-quality datasets for training intelligent systems for identifying and classifying cancer tissue, which makes algorithm evaluation largely subjective or reliable only for testing with minimal reliability.

The challenges described above are not unique to the field of histopathology images analysis systems. CNN and ViT have different toolsets to approach them.

Convolutional Neural Networks (CNNs) are a class of deep learning models that have significantly advanced the field of machine vision by enabling automatic learning of hierarchical representations from image data [8].

The convolution operation in position (x, y) is defined by formula 1.

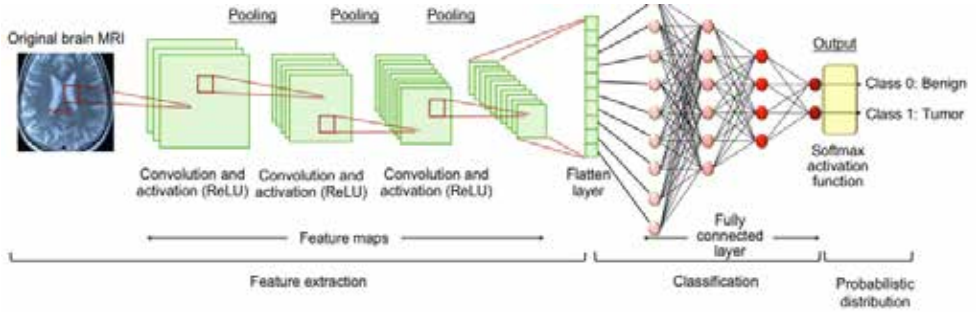
$$(I * K)(x, y) = \sum_{i=0}^{a} \sum_{j=0}^{b} I(x + i, y + j) \cdot K(i, j) \quad (1)$$

where I represents the input image, K denotes the kernel (or filter), a and b are the dimensions of the kernel K, (x, y) is the position in the output feature map, (i, j) is the position in the kernel K, and * symbolizes the 2D convolution operation. The convolution operation is performed by sliding the kernel K over the input image I and computing the dot product between the kernel and the corresponding region of the image at each position (x, y). The result is a new feature map where each pixel value is the sum of the element-wise products between the kernel and the corresponding image region.

Convolution applied of original image pixel values with kernel also called a receptive field. Convolution is applied with fixed stride size over the whole image. In such a way it forms a feature map. There could be several feature maps at once – each characterized by a different kernel. The feature map in a set represents some specific feature. Upper layers feature maps can be trained to detect low level features like edges and textures. Deeper layers detect more complicated features – for instance the presence of nucleus of specific type. Usually, convolution layers are interleaved with pooling layers – which provide the property of local translation invariance. Last layer output contains internal representation – learned features in the process of machine learning, these values are feeded to the classification layer (also called head), which provides the end probabilistic distribution on recognition classes, see picture 1.

There are several features of CNN which leverage its high efficiency:

- Parameter sharing – the weights of the kernel are shared while applying to different parts of the image. This property means that there are less parameters to learn – in such a way CNN has better performance, and less memory consumption compared with multilayer perceptron (MLP). Parameter sharing also leverages prevention of overfitting [9].
- CNN networks are characterized by translation equivariance – thus demanding less training examples to generalize well.



Pic. 1. Topology of typical CNN [7]

Despite described advantages of this architecture, CNNs are known to struggle with understanding relationships between distant parts of an image because their receptive field size is limited. Their fixed kernel sizes prevent them from recognizing features at varying scales and in diverse contexts across the entire image. Partially this is mitigated by the fact that deeper layers of the network have receptive fields transitively covering a bigger area of initial image. However, this architecture implies loss of information while transitioning between layers. To mitigate the loss of information problem – was implemented approaches like residual connections [10].

A relatively new approach includes application of transformer architecture in computer vision – visual transformers (ViT). Initially an approach was proposed in work [11]. It is quickly becoming the dominant architecture in tasks of computer vision – having outstanding accuracy and operation time [12].

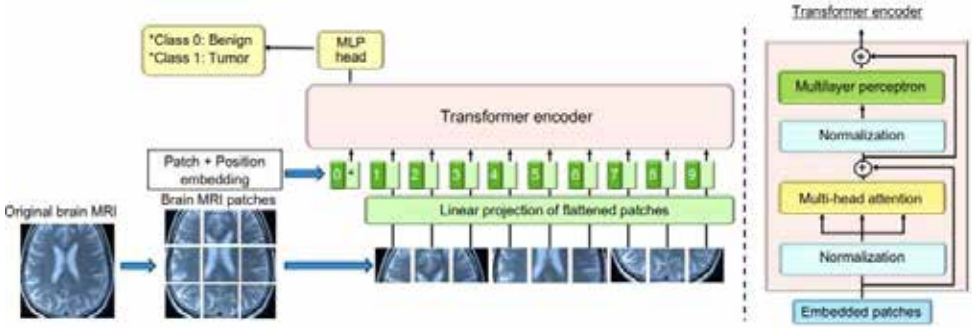
The core operation in transformer architecture is attention applied between embeddings of input sequence, formula (2)

$$\text{Attention}\left(Q,K,V\right)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{2}$$

where Q, K, and V are the query, key, and value matrices respectively, QK^T represents the matrix multiplication of Q and the transpose of K, d_k is the dimensionality of the key vectors used as a scaling factor, and the softmax function is applied to the scaled dot product to obtain the attention weights. The attention operation computes a weighted sum of the value vectors, where the weights are obtained by the compatibility of the query with the corresponding keys. In case of ViT – the initial image is cut on patches, each patch is represented by embedding, see image 2.

The advantages of ViT include ability to capture long range dependency between different areas of the image while forming the decision rule. Also, training of this kind of architecture can be effectively scaled by the process of parallelization, compared to long short-term memory networks (LSTM) architecture, where this is inconvenient or impossible to achieve [13].

At the same time ViT architecture is missing some of the basic inductive biases which CNN have. For example, translation equivariance and locality. However, it seems like larger datasets (14M-300M images) compensate for the lack of these biases and transformer architecture outperforms CNN based solutions [11]. In many fields like histopathology analysis, it is almost impossible to prepare the dataset of such size. For this reason, it is usually used in conjunction with transfer learning techniques [14].



Pic. 2. ViT architecture [7]

IET technology is well known to be used together with CNN based automatic feature extraction [15]. However, transformer architecture as a backbone and IET as classifier have not been researched enough. Having in mind the high relevance of the problem, the goal of the paper is to develop an intellectual system with ViT based neural network as the backbone and IET for classification.

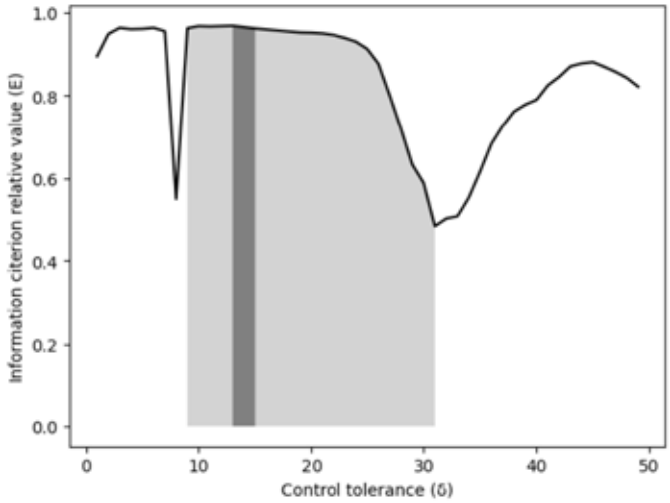
Development of the system. Consider the task of binary classification in context of the task of breast cancer histopathological imaging analysis. Let $\left\{ \boldsymbol{X}_{\boldsymbol{m}}^{0} \mid \boldsymbol{m} = \overline{1, \boldsymbol{M}} \right\}$ be the alphabet of recognition classes where M – number of recognition classes in our case 2. $\boldsymbol{X}_{1}^{0}$ – is a class of benign tissue type, while $\boldsymbol{X}_{2}^{0}$ – is a class of malignant tissue type. The backbone refers to the feature-extracting network that processes input data into a certain object property training matrix $\left\| \boldsymbol{y}_{\boldsymbol{m},\boldsymbol{i}}^{(j)} \right\|, \boldsymbol{i} = \overline{1, \boldsymbol{N}}, \boldsymbol{j} = \overline{1, \boldsymbol{n}}$, where N, n are the number of classes features recognition and implementations, respectively. In the process of machine learning having as input training matrix IET algorithm is seeking vector of optimal parameters $\boldsymbol{g}_{\boldsymbol{m}} = \boldsymbol{x}_{\boldsymbol{m}}, \boldsymbol{d}_{\boldsymbol{m}}, {}'_{k}$, by maximizing the averaged information criterion $\bar{\boldsymbol{E}} = \frac{1}{\boldsymbol{M}} \sum_{\boldsymbol{m}=1}^{\boldsymbol{M}} \max_{\boldsymbol{G}_{\boldsymbol{E}} \cap \{\boldsymbol{k}\}} \boldsymbol{E}_{\boldsymbol{m}}^{(\boldsymbol{k})}$. As information criterion was used modified Kulback measure:

$$\mathrm{E}_{m}^{(k)} = \frac{n - \left(K_{1,m}^{(k)} + K_{2,m}^{(k)} \right)}{n} \log_2 \frac{2\mathrm{n} + \xi - \mathrm{K}_{1,\mathrm{m}}^{(k)} - \mathrm{K}_{2,\mathrm{m}}^{(k)}}{\mathrm{K}_{1,\mathrm{m}}^{(k)} + \mathrm{K}_{2,\mathrm{m}}^{(k)} + \xi} \tag{3}$$

As the backbone was used Swin v2 transformer [16]. Swin Transformer V2 (SwinV2) offers several advantages over Vision Transformer (ViT), particularly in terms of scalability, efficiency, and performance across various vision tasks. Swin v2 applies the attention operation in fixed windows – thus the complexity of the overall attention routine is proportional to the number of windows in image – compared to quadratic complexion with global attention. To attain details across the whole image – implemented window shifting process, during which adjacent windows information is combined – thus formed hierarchical feature maps. This peculiarity of the model allows it to have over three billion parameters. Model takes as input an RGB image having 224×224 resolution. The patch size is 4×4, embedding 192, window size 7×7 patches. The input image is passed via a series of transformer blocks, each transformer block consists of layer normalization layer, scaled cosine attention across patches in the window. The output of the feature extraction stage has dimensions of 1536×8×8. As the final step we applied adaptive average pooling – thus obtaining a single vector size of 1536 representing the

whole image. For our purposes a large model type containing 3 billion parameters was used, pretrained on Imagenet-22K dataset [17]. Model fine tuning was organized on the BreakHis dataset [18], the dataset was splitted in proportion: 70% for training and 30% for evaluation. The following hyperparameters were used: learning rate 5*e*-05, training batch size 8, total number of epochs 7.

The pretrained Swin v2 backbone was used to prepare the training matrix for IET classifier. Each realization vector consists of 1536 features scaled to range [0, 100]. Training was conducted in parallel mode – optimal parameters for each feature were calculated simultaneously. Selection level was chosen to be 0.5 while max value of control tolerances was chosen to be half of the range of feature values – 50. The graph of normalized information criterion by control tolerance is presented on picture 3.



Pic. 3. Graph of normalized Kulback information criterion by control tolerance (light gray is working area, dark gray is area of global maximum of information criterion)

The normalized information criterion reached a maximum value of 0.97 (out of maximum 1), on control tolerance value of 12.

Results. For evaluation of the functional efficiency of the model the remaining 30% of the BreakHis dataset was used. To have a balanced test set, 847 samples for benign and malignant class each. In the exam mode the input image was feeded into backbone – obtaining a vector representing the image. After that, the vector was passed to the IET classifier to apply the set of decisive rules. The accuracy of the classifier reached 98.5%. Compared with ResNet based backbone accuracy 88.98%, the new approach gives a substantial boost.

Conclusions. The paper compared the functional efficiency of two types of backbones applied in conjunction with IET for the binary classification of breast cancer. In the first case was used CNN backbone based on ResNet architecture in second case was used ViT backbone based on Swinv2. The test has shown that the accuracy of transformer-based backbone Swin v2 has almost 10% higher accuracy. Even though transformer-based architectures lack inductive biases which CNN based architectures have, having pre-training on big datasets eliminates this fact. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) exhibit different biases in image processing

tasks, with CNNs showing a strong inclination towards texture and ViTs demonstrating a moderate preference for shape. This distinction arises from the inherent architectural differences and the way these models process visual information. CNNs, due to their convolutional nature, tend to focus on localized texture patterns, while ViTs, which rely on self-attention mechanisms, are more adept at capturing global shape information [19]. It can be assumed that in the task of breast cancer analysis by histopathological images the shape of the nucleus has crucial importance. The prospective direction of the research in this area is development of ensembles of CNN and ViT models – such systems potentially can effectively use texture and shape information, in such a way producing robust decisive rules with high accuracy.

REFERENCES:

1. Aljehani, M. R., Alamri, F. H., K Elyas, M. E., Almohammadi, A. S., Alanazi, A. S. A., & Alharbi, M. A. (2023). The importance of histopathological evaluation in cancer diagnosis and treatment. *International Journal of Health Sciences*. https://doi.org/10.53730/ijhs.v7ns1.15270

2. Hosseini, M. S., Ehteshami Bejnordi, B., Trinh, V. Q.-H., Chan, L., Hasan, D., Li, X., Yang, S., Kim, T., Zhang, H., Wu, T., Chinniah, K., Maghsoudlou, S., Zhang, R., Zhu, J., Khaki, S., Buin, A., Chaji, F., Salehi, A., Nguyen, B. N., Samaras, D., & Plataniotis, K. N. (2024). Computational pathology: A survey review and the way forward. *Journal of Pathology Informatics*, 15, Article 100357. https://doi.org/10.1016/j.jpi.2023.100357

3. Greeley, C., Holder, L. B., Nilsson, E., & Skinner, M. K. (2024). Scalable deep learning artificial intelligence histopathology slide analysis and validation. *Dental Science Reports*, 14(1). https://doi.org/10.1038/s41598-024-76807-x

4. Laxmisagar, H. S., & Hanumantharaju, M. C. (2020). A survey on automated detection of breast cancer based histopathology images. In Proceedings of the 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 19-24). IEEE. https://doi.org/10.1109/ICIMIA48430.2020.9074915

5. He, L., Long, L. R., Antani, S., & Thoma, G. R. (2012). Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, 107(3), 538-556.

6. Stenkvist, B., Westman-Naeser, S., Holmquist, J., Nordin, B., Bengtsson, E., Vegelius, J., Eriksson, O., & Fox, C. H. (1978). Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations. *Cancer Research*, 38(12), 4688–4697.

7. Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M., & Hamamoto, R. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, 48(1), Article 84. https://doi.org/10.1007/s10916-024-02105-8

8. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. https://doi.org/10.1109/5.726791

9. Azadbakht, A., Kheradpisheh, S. R., Hassani, I. K., & Masquelier, T. (2022). Drastically reducing the number of trainable parameters in deep CNNs by inter-layer kernel-sharing. *arXiv preprint arXiv:2210.14151*.

10. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). IEEE. https://doi.org/10.1109/CVPR.2016.90

11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., &

Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=YicbFdNTTy

12. Wu, C., & He, T. (2024). A survey of applications of vision transformer and its variants. In *Proceedings of the 10th IEEE International Conference on Intelligent Data and Security (IDS)* (pp. 21-25). IEEE. https://doi.org/10.1109/IDS62739.2024.00011

13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)* (pp. 6000-6010).

14. Lee, R. S. T. (2024). Transfer learning and transformer technology. In *Natural Language Processing: A Textbook with Python Implementation* (pp. 175-197). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-1999-4_8

15. Papchenko, O., Kuzikov, B. (2025). Hybrid deep learning and information-extreme approach for breast cancer histopathological image classification. *Herald of Khmelnytskyi National University. Technical Sciences*, Vol. 347 (Issue 1), pp. 175-181. https://doi.org/10.31891/2307-5732-2025-347-24

16. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., & Guo, B. (2021). Swin transformer V2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*. https://doi.org/10.48550/arXiv.2111.09883

17. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211-252. https://doi.org/10.1007/s11263-015-0816-y

18. Spanhol, F., Oliveira, L., Petitjean, C., & Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, 63, 1455-1462. https://doi.org/10.1109/TBME.2015.2496264.

19. Iwata, A., & Okuda, M. (2024). Quantifying Shape and Texture Biases for Enhancing Transfer Learning in Convolutional Neural Networks. *Signals*, 5(4), 721-735. https://doi.org/10.3390/signals5040040