Journal of Computer and Communications, 2023, 11, 1-18
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Simplified Inception Module Based Hadamard Attention Mechanism for Medical Image Classification

Yanlin Jin[1], Zhiming You[2*], Ningyin Cai[1]

[1]Department of Statistics, Guangzhou University, Guangzhou, China
[2]Department of Economics, Jinan University, Guangzhou, China
Email: *bestyjimmy@foxmail.com

How to cite this paper: Jin, Y.L., You, Z.M. and Cai, N.Y. (2023) Simplified Inception Module Based Hadamard Attention Mechanism for Medical Image Classification. *Journal of Computer and Communications*, **11**, 1-18.
https://doi.org/10.4236/jcc.2023.116001

Received: May 7, 2023
Accepted: June 9, 2023
Published: June 12, 2023

Abstract

Medical image classification has played an important role in the medical field, and the related method based on deep learning has become an important and powerful technique in medical image classification. In this article, we propose a simplified inception module based Hadamard attention (SI + HA) mechanism for medical image classification. Specifically, we propose a new attention mechanism: Hadamard attention mechanism. It improves the accuracy of medical image classification without greatly increasing the complexity of the model. Meanwhile, we adopt a simplified inception module to improve the utilization of parameters. We use two medical image datasets to prove the superiority of our proposed method. In the BreakHis dataset, the AUCs of our method can reach 98.74%, 98.38%, 98.61% and 97.67% under the magnification factors of 40×, 100×, 200× and 400×, respectively. The accuracies can reach 95.67%, 94.17%, 94.53% and 94.12% under the magnification factors of 40×, 100×, 200× and 400×, respectively. In the KIMIA Path 960 dataset, the AUCs and accuracy of our method can reach 99.91% and 99.03%. It is superior to the currently popular methods and can significantly improve the effectiveness of medical image classification.

Keywords

Deep Learning, Medical Image Classification, Attention Mechanism, Inception Module

1. Introduction

With the continuous innovation of technology and the development of medical image classification, the field of clinical medicine has ushered in a new change.

Through medical image technology, modern medicine can observe the details of the internal tissue structure of the human body, and suit the remedy to the specific case, which can better help patients recover. As Medical imaging technology is widely used in the field of clinical medicine, effective classification technique of medical images has played a vital role in assisting clinical treatment and nursing. For the classification of medical images, the traditional machine learning method was initially used, but it is still not ideal in classification performance, because the technical innovation is slow, and it takes a lot of time to process large-scale data.

After 2012, the methods based on deep learning began to be widely used in the field of medical image classification. Nowadays, the medical image classification technology has made great progress. Yadav *et al.* [1] made a summary of CNN used in medical image classification technology, and reached the conclusion that using deep CNN is more effective for medical image classification. Jeyaraj *et al.* [2] used CNN to classify oral cancer, with an accuracy of 91% and reached the same conclusion. Wang *et al.* [3] used the deep convolution neural network to classify medical images and also achieved good results. However, many parameters in the deep convolution neural network are not used, and the utilization rate of parameters is low. In order to improve the parameter utilization, Zou *et al.* [4] used the convolution neural network of inception V1, and compared the effect of spatial pyramid pooling and global average pooling. They found that global average pooling was more effective than spatial pyramid pooling. Lu *et al.* [5] proposed an image classification method based on inception net for classification of skin cancer images, with an accuracy of 100%, but it needs a large number of datasets for training. Mahin *et al.* [6] used inception V3 and transfer learning methods to classify COVID-19, and obtained satisfactory results. However, the experimental process is complex and requires a lot of time. The above methods are all used in the classification of medical images by CNN and have achieved certain results, but the experiments have their own shortcomings. The use of deep learning methods for medical image classification often yields good results due to the strong fitting ability of neural networks. However, it also increases calculation complexity, which leads to higher equipment requirements and time consumption for deep learning methods for medical image classification. In clinical medicine, time-saving and accurate medical image classification methods can help patients improve the success of treatment. Therefore, it is necessary to propose an accurate and efficient deep-learning method for medical image classification.

In recent years, attention mechanism [7] has been proposed and widely used to deal with neural network problems. In the field of medical imaging, attention mechanism has been used to deal with the problem of medical image classification. Xu *et al.* [8] proposed a CNN global spatial attention mechanism for medical image classification, and achieved good results. But due to his proposed method utilizing the pixels of the image, the effectiveness will be weakened when the image quality is low. Zou *et al.* [9] embedded attention mechanism and high-

order statistical representation into residual convolution neural network when processing pathological images of breast cancer, and obtained 85% of the best patient level classification accuracy in Bach database. Ning *et al.* [10] Proposed a method of skull fracture image classification based on attention, gave different weights to the feature information extracted by Resnet to get good results. Some predecessors also used the channel attention mechanism to classify medical images. The method has too many parameters, resulting in insufficient parameter utilization. Wang *et al.* [11] proposed a residual network model based on the channel attention mechanism, the average F1 score is 88.2%. However, this method greatly increases the calculation complexity and requires more time. Du *et al.* [12] proposed a new efficient channel attention depth dense convolutional neural network, and the classification accuracy is as high as 90%. This model uses a deep convolutional model, when encountering problems such as gradient vanishing or gradient explosion, it will have a significant impact on the result. The above results show that the combination of attention mechanism and neural network can greatly improve the performance of the model, but these methods require a large amount of calculation and have certain limitations, there is room for further improvement.

To solve the above-mentioned problem, we propose a simplified inception module based Hadamard attention mechanism.

The contribution of our work is as follows:

1) We propose a new attention mechanism: Hadamard attention mechanism, it makes important information more important and suppresses invalid information by using Hadamard product. It improves the accuracy of medical image classification without greatly increasing the complexity of the model;

2) The traditional inception model is improved, and a simplified inception model is proposed to reduce the number of model parameters, improve the utilization of parameters, and speed up the efficiency of model operation;

3) In order to further improve the effectiveness of medical image classification, Hadamard attention mechanism and inception V1 are combined to build a new GoogleNet architecture, which greatly improves the accuracy of medical image classification and makes the results more convincing (The following paragraph has been deleted).

2. Materials and Methods

2.1. Overall Framework

In this section, we illustrate the architecture of our proposed model named Hadamard attention mechanism based simplified inception model. The whole model is depicted in **Figure 1**. Compared with the previous GoogleNet [13], We simplified the structure of the original inception V1 module, decrease the number of inception module from 9 to 6 and replace all Local Response Normalization (LRN) [14] with Batch Normalization (BN) [15]. Our model starts with a simple but effective CNN model with 7 different layers aiming to extract features

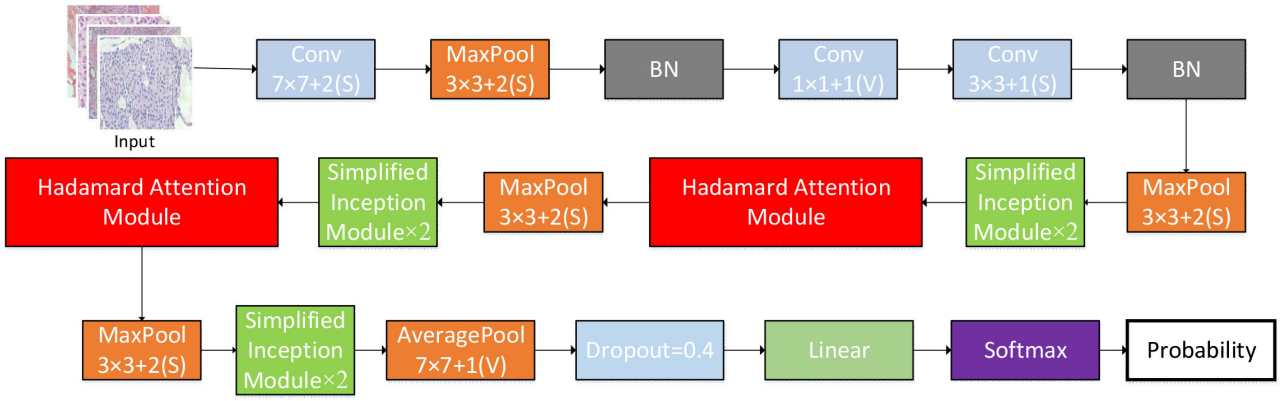

Figure 1. The overall framework of simplified inception is based on Hadamard attention mechanism. Network mainly consist of 6 simplified inception module and 2 Hadamard attention mechanism.

primarily. Then 6 simplified inception modules (SI) are attached to end of the CNN above with one Maxpool layer between every two SIs. These SIs work as multi-scale feature extractors as same as the inception modules in GoogLeNet. Furthermore, we proposed a brand new attention mechanism called Hadamard attention (HA) mechanism which is simple to implement and effective. The following experiment results show that our HA incorporation with SIs are beyond many popular deep learning models.

2.2. Hadamard Attention Mechanism

Attention mechanism has been widely used in computer vision and natural language processing. Attention mechanism is mainly divided into spatial attention, channel wise attention and mix attention mechanism. Its representatives mainly include: Self-Attention Mechanism (SAM) [16], Squeeze-and-Excitation Networks (SE-Net) [17] and Convolution Block Attention Module (CBAM) [18]. Their common point is that they all require a large number of convolutions and product operations which may lead to overfitting and high computational complexity poor performance when facing small datasets. In particular, because of the need to learn global features, SAM needs a feature map of appropriate size as input. If the feature map is too large, it is difficult for SAM to fully extract useful information; if the feature map is too small, it is close to the classification decision-making level, the wrong learning results of SAM can easily have a negative impact on backbone model. To address the issues above and inspired by the Cross-Attention Networks [19], we proposed a brand new attention mechanism called Hadamard Attention mechanism, the specific structure is shown in **Figure 2**.

In this module, we regard the size of the feature map element as the weight of the feature at the corresponding position. Firstly, two 1×1 convolution kernels are used to map the input feature map respectively. This decision was based more on convenience rather than necessity. The purpose of 1×1 convolution kernel is to amplify the weight of important information and reduce the weight

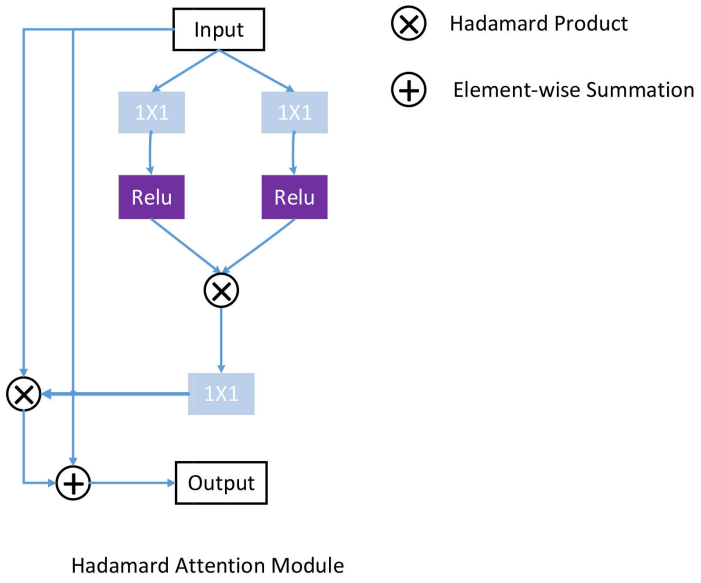

Figure 2. Hadamard attention mechanism model diagram.

of invalid information. It is worth noting that the 1 × 1 convolution kernel here not only maps the input feature map, but also reduces the number of channels of the feature map, so as to reduce the computational complexity of subsequent operations. And then, in this article, two ReLu activation functions are used to filter out the non-pixel information generated in the convolution process. In order to further increase the weight of important information, Hadamard product is applied to two filtered results. Then, a 1 × 1 convolution kernel is used for feature map mapping, and the channel number of the feature map after mapping is restored to be consistent with the original input. Finally, as the same with the residual attention network [20], the final output is obtained by adding the residual learning module and Hadamard product operation. If X is the input of HA, the output of HA can be expressed as:

$$\mathrm{HA}(X) = (1 \oplus F(X)) \otimes X \tag{1}$$

In the above formula, F represents the above series of convolution and product operations, $\oplus$ is element-wise summary and $\otimes$ is Hadamard product. It can be seen from equation 1 that learns the weight of each element from it, and thus enhances the important information in it by means of residual learning and Hadamard product, while the invalid information is suppressed.

1 × 1 convolution kernel is the key to HA. On the one hand, like SE -Net, 1 × 1 convolution kernel first reduces the number of channels of feature map, and then restores them to the original number of channels; But SE-Net only learns the weights of different channels, and the number of channels that are too large or too small will affect the learning ability of SE-Net. In our proposed Hadamard attention mechanism, it is not necessary to explicitly learn the weights of each channel, but directly use 1 × 1 convolution kernel to fuse elements at the same position of each channel, the parameters of the 1 × 1 convolution kernel are ac-

tually the weights of those elements. This makes our mechanism more flexible. On the other hand, 1 × 1 convolution kernel can effectively reduce the computational complexity. If the size of the feature map is $n \times d$ and the number of channels is c, the computational complexity of the self-attention mechanism is $O(n^2dc)$, while the computational complexity of the proposed HA is $O(ndc)$, which reduces the computational complexity by an order of magnitude.

2.3. Simplified Inception Module

Inception V1 is the main module of GoogleNet. Inception V1 mainly improves the convolution layer in the network, and uses multiple convolution kernels of different sizes to perform parallel convolution operations, trying to approximate large-scale sparse structures with multiple dense filters. There have been many improvements to inception V1, such as inception V2, V3 [21] and V4 [22], whose main purpose is to reduce the number of parameters and prevent overfitting. But these models still have the disadvantage of heavy burden for training. The experimental part in the next chapter shows that GoogleNet performs poorly on small datasets and converges slowly on large datasets. To solve the above problems, we propose the simplified inception module. The specific structure is shown in **Figure 3**. The module first reduces the original four 1 × 1 convolution kernel to one, which can greatly reduce the number of parameters. Secondly, in order to further reduce the number of parameters, all 5 × 5 convolution kernels are decomposed into two sequential 3 × 3 convolution kernels. As Szegedy mentioned in [21], this can not only reduce the number of parameters, but also keep the receptive field of 3 × 3 convolution consistent with that of 5 × 5 convolution. Imitating the dilated perception network [23], we can also integrate these two 3 × 3 convolution kernels to hole convolution to further expand the receptive field of SI, but this is not necessary. In order to accelerate the convergence speed of the model and solve the problem of internal covariate shift, we add a layer of BN layer after depthconcat. The traditional ReLu activation function is easy to generate dead cells, which leads to the stagnation of network learning. We use leakyrelu activation function [24] to replace ReLu.

We also consider the general design principles of the network [10]: with the increasing depth of network, the size of feature map should gradually decrease and the number of channels should gradually increase. This can help the network to speed up the training without losing too much information each time the size of the feature map is reduced. **Table 1** lists the size and channel number

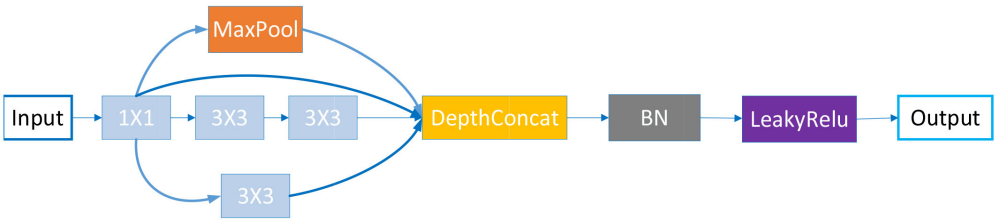

Figure 3. Simplified inception module model diagram.

Table 1. The size and channel number of output feature map of each layer in the simplified inception model.

type	patch size/stride	output size	1 × 1	3 × 3	two sequential 3 × 3	maxpool
convolution	7 × 7/2	112 × 112 × 32				
max pool	3 × 3/2	56 × 56 × 32				
batch norm		56 × 56 × 32				
convolution	3 × 3/1	56 × 56 × 128	64			
batch norm		56 × 56 × 128				
max pool	3 × 3/2	28 × 28 × 128				
SI 1		28 × 28 × 256	64	96	32	64
SI 2		28 × 28 × 416	112	128	64	112
max pool	3 × 3/2	14 × 14 × 416				
SI 3		14 × 14 × 576	160	160	96	160
SI 4		14 × 14 × 740	208	196	128	208
max pool	3 × 3/2	7 × 7 × 740				
SI 5		7 × 7 × 900	256	228	160	256
SI 6		7 × 7 × 1010	304	206	196	304
average pool	7 × 7/1	1 × 1 × 1010				
dropout (40%)		1 × 1 × 1010				
linear		1 × 1 × class num				
softmax		1 × 1 × class num				

of output feature map of each layer in the model. It can be seen that the models proposed by us all follow the above principles. "1 × 1", "3 × 3", "two sequential 3 × 3" and "max pool" stand for the number of channels output by corresponding kernels or max pool.

The key of SI is also 1 × 1 convolution kernel. If this 1 × 1 convolution kernel does not reduce the number of channels too much, and 3 × 3 and 5 × 5 convolution kernels can be gradually increased, the performance of SI will not be weakened or even improved. In a word, the improvement of SI is mainly to reduce the number of parameters and accelerate the convergence speed of the network, but also to make the model performance not reduce or even improve. Subsequent experiments show that the effect of the simplified perception module is not worse than that of inception v1.

3. Results and Discussion

3.1. Dataset

In this article, we use two different datasets. The basic information of these two datasets is shown below.

- BreakHis [25]: BreakHis dataset has computed tomography images of benign and malignant breast cancer at different magnification factors (40×, 100×, 200× and 400×). In addition, the dataset has the characteristic of category imbalance, including 2480 benign and 5429 malignant samples. We made a data preprocessing pipeline for augmenting and transforming images to be used during training. Specifically, the pipeline consists of the following steps: First, we resized the image to a specified size, which is 224 × 224 in this case. Second, for each sample or image in the dataset, we randomly applied a set of augmentation operations, including Gaussian blur, random vertical flip, random horizontal flip, and color jitter. The probability of applying each operation is 0.5. Finally, we converted the image data to the PyTorch tensor format and normalizes the pixel values to the range of [0, 1]. **Table 2** shows the specific information. In the next experiments, we train the model using samples with different magnification factors respectively. The images are publicly available at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
- KIMIA Path 960 [26]: KIMIA Path 960 has 20 classes of whole slide images of different kinds of tissue such as muscle, epithelial and so on. Each class has the same size of samples which is 48. Hence, the dataset has a total of 960 images and no class-imbalance exists, Data preprocessing is consistent with Breakhis. The images are publicly available at https://kimialab.uwaterloo.ca/kimia/index.php/pathology-images-kimia-path960/

3.2. Evaluation Measures

In this article, we first measure and analyze the classification ability of the proposed model and then we analyze the inference time and throughput of proposed model, comparing it with many other popular and classical models.

In this article, accuracy and AUC are used to measure and analyze the performance of the proposed model. Accuracy measures the ability of a model to classify all samples. In the binary-label classification task, the formula for accuracy is shown below.

Table 2. Specific information of BreakHis.

Magnification Factors	Benign	Malignant	Total
40×	652	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total of images	2480	5429	7909

$$\text{accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{2}$$

where TN is true negatives, TP is true positives, FN is false negatives and FP is false positives. As to multi-label classification task, top-k accuracy is adopted instead of accuracy used in binary-label classification task, where k denotes k classes of k highest probability. In one specific decision tensor, if the label corresponding to the k classes of k highest probability includes the only one ground truth label, then the prediction can be considered as a success. Compared to the accuracy, top-k accuracy is more flexible.

However, accuracy does not properly measure the performance of the model when facing the category imbalanced data, so we introduce the AUC metric. The AUC is Area below ROC curve. The ROC is a curve based on the values of TPR and FPR at different classification thresholds. The equations for TPR and FPR are shown below.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4}$$

Inference time measures the time it takes for the model to process a single sample. To measure the inference time properly, we create 300 different dummy inputs with batch sizes equal to 1 and calculate the average time for processing those samples sequentially. Throughput is the number of samples that the model can process in parallel per second. Unlike the inference time, we create 300 different dummy inputs with the largest batch size that our used device can process once and record the total spent time for process those samples. Thus, throughput is calculated by the following equation:

$$\text{throughput} = \frac{\text{batch size} \times 300}{\text{total time}} \tag{5}$$

3.3. Loss

As mentioned in Section 3.1, we use the BreakHis dataset with the category imbalance property. It can be noticed that if the cross-entropy function is used as the loss function for model training, not only the final accuracy is lower, but also the model tends to reach the bottleneck prematurely. For example, the loss keeps decreasing, but the accuracy remains the same. The reason is that the cross-entropy function does not help learn information about those hard-to-classify samples whose output category probabilities always fluctuate between 0.4 and 0.6. But the probabilities of those easy-to-classify samples keep approaching 1 or 0. To solve this problem, we adopt Focal Loss function [27] for model training, which is a modified version of the cross-entropy function, with the following equation:

$$Loss\left(p_{i,j}, y_{i,j}\right) = \sum_{i=1}^{N} -\left(1 - p_{i,j}\right)^{\gamma} \times \log\left(p_{i,j}\right) \tag{6}$$

where $y_{i,j}$ denotes the ground truth label j of the i^{th} sample, $p_{i,j}$ represents the probability of the i^{th} sample predicted as j by model, and γ is a hyperparameter. In this article, we set $\gamma = 2$. The advantage of Focal Loss is that it reduces the loss of easy-to-classify samples and increases the loss of hard-to-classify samples.

3.4. Training Method

This article deploys the model on a Tesla P100 GPU for training. Based on the Pytorch, this article uses Adam algorithm as the optimization algorithm, where the initial learning rate is 0.00001. For the learning rate adjustment strategy, the WarmUp strategy [28] is used. The WarmUp strategy starts with a small learning rate, which increases as the number of training rounds increases until the specified number of training rounds is reached. Then the learning rate decreases with the number of learning rounds. The update strategy used in this article can be expressed by the following equation.

$$lr = 0.0001 \times \begin{cases} \log(e+1), & e < epoch_{warmup} \\ \exp(-e), & e \geq epoch_{warmup} \end{cases} \tag{7}$$

where $epoch_{warmup}$ is the number of model warm-up training rounds, we set $epoch_{warmup} = 50$, and the total number of training rounds is 200.

3.5. Results and Discussion

In this section, this article first illustrates the classification effectiveness of our proposed model compared with many other popular and classical models. And then, this article investigates the effect of the attention mechanism in the proposed model, comparing the performance of SI with and without CABM, SE-Net, SA and HA respectively. Finally, this article demonstrates the inference time and throughput of the proposed model.

3.5.1. Classification Ability

Table 3 and **Table 4** show the experimental results of the different models under different magnification factors. **Table 3** and **Table 4** show the experimental results of the different models under different magnification factors. This article compares the proposed model with popular and classical models, such as CNN [29], VGG16 [30], AlexNet [31], GoogleNet [13], inception v3 [21], Texture CNN [32], CSDCNN [33], and compares with state-of-the-art models, such as Optimised CNN [34], CNN-LSTM [35], C-Net [36]. Obviously, in terms of accuracy and AUC, our proposed model outperforms popular and classical models. When compared with state-of-the-art models, our model performs better than Optimized CNN on two datasets. When comparing with CNN-LSTM and C-Net, we found that our method achieved better results at certain magnification on the breakhis, while C-Net performed better on the path 960.

Multi-scale convolution-based models, such as SI, GoogleNet and Inception V3, outperform other types of models, both in the BreakHis dataset and the KIMIA Path 960 dataset. Multi-scale convolution allows features to be extracted

Table 3. The experiment was compared with other models under different magnifications in BreakHis dataset.

Method	Magnification Factors							
	40×		100×		200×		400×	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
SI + HA	0.9874 ± 0.0054	0.9567 ± 0.0119	0.9838 ± 0.0044	0.9417 ± 0.0134	0.9861 ± 0.0042	0.9453 ± 0.01	0.9767 ± 0.0037	0.9412 ± 0.0054
CNN	0.8127 ± 0.0018	0.8076 ± 0.0131	0.8359 ± 0.002	0.8570 ± 0.0054	0.8566 ± 0.0179	0.8859 ± 0.0112	0.8165 ± 0.0095	0.8247 ± 0.0153
VGG 16	0.8416 ± 0.002	0.8307 ± 0.0102	0.8030 ± 0.0148	0.8186 ± 0.0046	0.8221 ± 0.0022	0.8443 ± 0.0235	0.8056 ± 0.0134	0.8244 ± 0.0152
AlexNet	0.8033 ± 0.0054	0.8347 ± 0.0103	0.8062 ± 0.0027	0.8166 ± 0.0035	0.8674 ± 0.0088	0.8676 ± 0.017	0.8566 ± 0.0382	0.8554 ± 0.0021
Inception V3	0.9023 ± 0.0053	0.8742 ± 0.0123	0.8925 ± 0.0021	0.8633 ± 0.0133	0.8924 ± 0.0054	0.8717 ± 0.0067	0.8811 ± 0.0034	0.8653 ± 0.0079
Texture CNN	0.9123. ± 0.0077	0.8966 ± 0.0054	0.8837 ± 0.0044	0.8644 ± 0.021	0.8976 ± 0.0135	0.8744 ± 0.0089	0.9023 ± 0.0039	0.8831 ± 0.0107
CSDCNN	0.9795 ± 0.0036	0.9480 ± 0.0087	0.9766 ± 0.057	0.9324 ± 0.0105	0.9757 ± 0.0105	0.9355 ± 0.013	0.9741 ± 0.0044	0.9401 ± 0.0098
Optimised CNN	0.9322 ± 0.0046	0.9034 ± 0.0023	0.8977 ± 0.0043	0.8635 ± 0.0104	0.8614 ± 0.0103	0.8317 ± 0.0234	0.8388 ± 0.0045	0.8144 ± 0.0317
CNN-LSTM	0.9834 ± 0.0033	0.9553 ± 0.0027	0.9855 ± 0.0057	0.9465 ± 0.0112	0.9877 ± 0.0035	0.9501 ± 0.0173	0.9649 ± 0.0023	0.9318 ± 0.0044
C-Net	0.9933 ± 0.0015	0.9677 ± 0.0136	0.9967 ± 0.0017	0.9601 ± 0.0023	0.9769 ± 0.0024	0.9293 ± 0.0011	0.9755 ± 0.0026	0.9317 ± 0.0056

Table 4. The experiment compared with popular models in KIMIA Path 960.

Method	SI + HA	CNN	VGG 16	AlexNet	Inception V3	Texture CNN	CSDCNN
AUC	0.9991 ± 0.0003	0.8433 ± 0.0137	0.8703 ± 0.0216	0.8214 ± 0.0152	0.9051 ± 0.0061	0.9266 ± 0.0061	0.9951 ± 0.0005
Accuracy	0.9903 ± 0.0024	0.8824 ± 0.0045	0.8431 ± 0.0055	0.8724 ± 0.0053	0.8824 ± 0.0077	0.9034 ± 0.0164	0.9901 ± 0.0021

from images at different scales. Specifically, in medical image datasets, both benign and malignant tumors have features of different scale sizes, such as the size and color of cell nucleus, the shapes of cell, etc. Although CNN-based models can use different sizes of convolutional kernels in different layers, using only one type of convolutional kernel in each layer can only learn features at one scale

and there is no guarantee that all the learned features can be remembered up to the classification layer.

However, among the multi-scale based models such as SI, GoogleNet and Inception V3, SI outperforms the other models. Moreover, there are fewer parameters in SI compared to GoogleNet and Inception V3. In SI, since two 3 × 3 convolutional kernels are used instead of only one 5 × 5 convolutional kernel, then, in fact, the depth of SI is actually deeper than GoogleNet, inception V3. This is in line with what was demonstrated in [30]: for a given perceptual field, to some extent, increasing the network depth can improve the performance of the model. In addition to this, the reduction in the number of parameters for SI mainly comes from the reduced use of 1 × 1 convolution kernels. It is worth noting that the results of GoogleNet were obtained after 250 rounds of training, while the other models were trained for only 150 rounds. **Figure 4** and **Figure 5** show the curves of loss and accuracy of SI-based model and GoogleNet. Among them, GoogleNet uses two metrics, accuracy for Breakhis-40×, top-1 accuracy and top-3 accuracy for KIMIA Path 960, respectively, and is trained for 250 rounds.

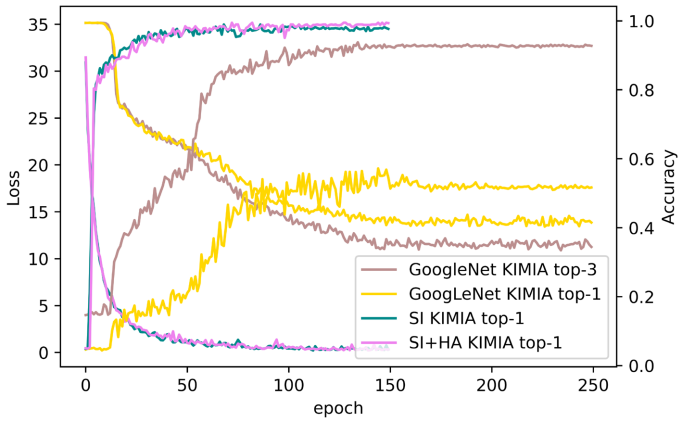

Figure 4. Curves of loss and accuracy of SI-based model and GoogleNet in KIMIA Path 960 dataset.

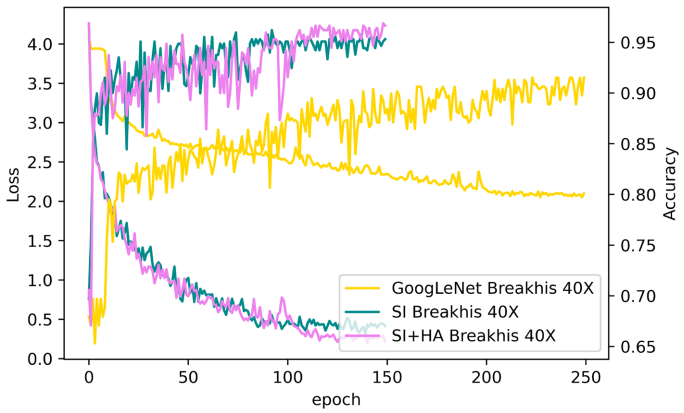

Figure 5. Curves of loss and accuracy of SI-based model and GoogleNet in BreakHis dataset.

As can be seen from Figure 4, GoogleNet not only converges slowly in the KIMIA Path 960 dataset, but also its top-1 accuracy only reaches about 50%, and top-3 accuracy barely exceeds 90%. In Breakhis-40×, the accuracy of GoogLeNet does not exceed 90 until around the 200th round. Therefore, it is feasible to replace the original four 1 × 1 convolutional kernels with only one 1 × 1 convolutional kernel, which not only can greatly reduce the number of model parameters without affecting the learning ability, but also helps to avoid overfitting and speed up the convergence of the model.

Table 5 and Table 6 show the comparison between the proposed HA mechanism and other attention mechanisms with SI as backbone. Although the HA mechanism does not always achieve optimal results for all samples of different magnification factors, it is not far from them. The overall performances of dual-attention mechanism CBAM based on the self-attention mechanism and the channel-attention mechanism are similar to the ones of HA, but HA has fewer parameters. As will be seen in later experimental sessions, the advantages of HA lie in the low inference time and high throughput.

Table 5. The experiment compared with state-of-the-art models in KIMIA Path 960.

Method	Optimised CNN	CNN-LSTM	C-Net
AUC	0.9514 ± 0.0065	0.9913 ± 0.0013	0.9977 ± 0.0005
Accuracy	0.9231 ± 0.0087	0.9877 ± 0.0017	0.9966 ± 0.0012

Table 6. Comparison experiments under different magnifications in BreakHis dataset.

Method	Magnification Factors							
	40×		100×		200×		400×	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
SI + HA	0.9874 ± 0.0054	0.9567 ± 0.0119	0.9838 ± 0.0044	0.9417 ± 0.0134	0.9861 ± 0.0042	0.9453 ± 0.01	0.9767 ± 0.0037	0.9412 ± 0.0054
SI	0.9923 ± 0.0029	0.9274 ± 0.0122	0.9712 ± 0.005	0.9111 ± 0.0127	0.9807 ± 0.0079	0.9249 ± 0.0192	0.9532 ± 0.0105	0.8923 ± 0.01
SI + SE	0.9778 ± 0.006	0.9321 ± 0.012	0.9776 ± 0.0048	0.9303 ± 0.0116	0.9837 ± 0.0072	0.9462 ± 0.0135	0.9822 ± 0.0034	0.9252 ± 0.0057
SI + SA	0.9735 ± 0.0086	0.9312 ± 0.0151	0.9575 ± 0.0083	0.9175 ± 0.0086	0.9767 ± 0.0063	0.9369 ± 0.0117	0.9894 ± 0.038	0.9557 ± 0.0029
SI + SA + SE	0.9874 ± 0.0057	0.9624 ± 0.0085	0.9791 ± 0.0059	0.9274 ± 0.0122	0.9903 ± 0.0033	0.9558 ± 0.0084	0.9704 ± 0.0052	0.9253 ± 0.0057

The advantage of HA is also located in the ability to maintain more stability while achieving excellent results. Figure 6 and Figure 7 shows the fluctuation of accuracy and AUC of SI-based models with different attention mechanism trained in 5 datasets for 50 times. It can be seen that SI + HA in general is very stable. On the contrary, SI + SA relatively speaking has a more highlighted instability.

And HA is well adapted to small samples dataset. In KIMIA Path 960, HA is the only attention mechanism that can improve the performance of the model. This is also not available in other large models. Other attention mechanisms, due to the introduction of too many parameters, instead weaken the learning ability of the model. (Table 7)

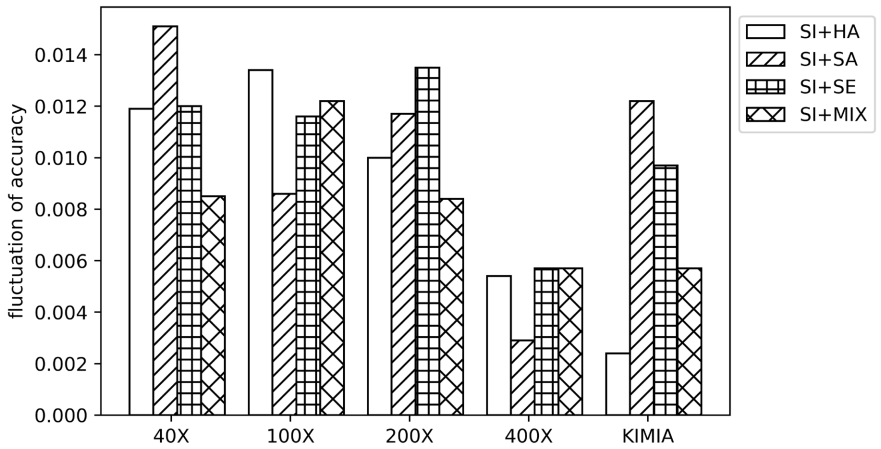

Figure 6. Fluctuation of accuracy on 5 different datasets.

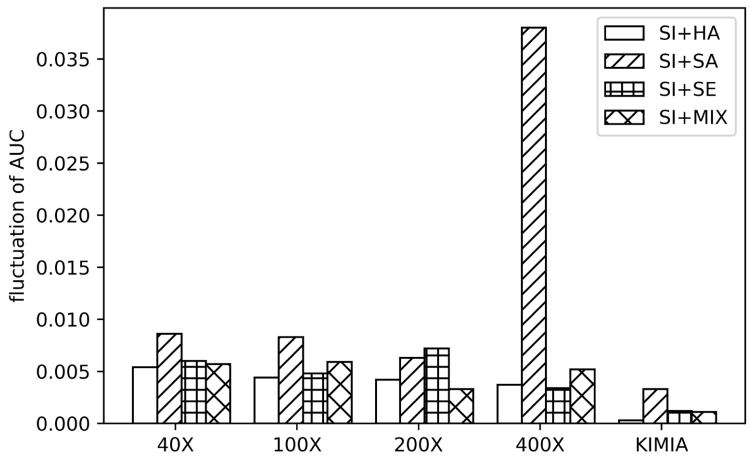

Figure 7. Fluctuation of AUC on 5 different datasets.

Table 7. Comparison experiments in KIMIA Path 960 dataset.

Method	SI + HA	SI	SI + SE	SI + SA	SI + SA + SE
AUC	0.9991 ± 0.0003	0.9983 ± 0.0005	0.9921 ± 0.0012	0.9913 ± 0.0033	0.9954 ± 0.0011
Accuracy	0.9903 ± 0.0024	0.9823 ± 0.0031	0.9796 ± 0.0097	0.9721 ± 0.0122	0.9824 ± 0.0057

Table 8. Inference time and throughput of different models.

Model	inference time(ms)	throughput
SI	3.8157	2417.6568
SI + HA	4.3107	1929.5900
SI + SE	4.3295	2266.5443
SI + SA	5.1588	1876.5547
SI + MIX	4.6068	1342.1604
GoogLeNet	7.2448	270.3429
VGG	12.4864	135.2229
AlexNet	10.6465	156.6873

3.5.2. Inference Time and Throughput

Table 8 illustrated the inference time and throughput of different models respectively. Not surprisingly, the inference time of SI is the smallest, while GoogLeNet's inference time is nearly twice that of SI. With the addition of the attention mechanism, HA and SE increase the inference time the least, both by about 0.5 ms each. And SA and MIX add 1.3 ms and 0.8 ms respectively. In terms of throughput, HA and SE are similar, while the throughput of the other mechanisms declines severely.

4. Conclusions

In recent years, medical image classification technology is more and more widely used in clinical medicine. In order to avoid misjudgment of the disease, clinical medicine has higher and higher requirements for the accuracy of image classification, this article proposes a simplified inception module based Hadamard attention. As an attention mechanism, Hadamard attention can give greater weight to key information and reduce computational complexity. If it is necessary to pay more attention to key information in medical image classification tasks in the future and do not want to increase computational complexity, Hadamard attention can be considered. Simplified inception module is the result of simplifying the inception v1, which greatly reduces the number of parameters while accuracy does not decrease. If the inception v1 network is used in medical image classification in the future, simplified inception module can be considered as a substitute for inception v1. However, SI has certain limitations. Although it can converge faster than inception v1, it may suffer from oscillation of convergence curve.

The experimental results show that the medical image classification accuracies and AUC of our models are higher than the existing popular models. However, there is still room for improvement when the proposed method is applied to the binary classification problem, and the accuracy of image classification can be improved through further optimization of the model. In addition, our proposed method can not only be used for medical image classification, but also can be

applied to other image classification problems, but it requires further attempts and improvements. we hope the above problem can be solved in the future work.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] Yadav, S., Rathod, R., Pawar, S.R., Pawar, V.S. and More, S. (2021) Application of Deep convulational Neural Network in Medical Image Classification. 2021 *International Conference on Emerging Smart Computing and Informatics* (*ESCI*), Pune, 5-7 March 2021, 120-129. https://doi.org/10.1109/ESCI50559.2021.9396854

[2] Jeyaraj, P.R. and Samuel Nadar, E.R. (2019) Computer-Assisted Medical Image Classification for Early Diagnosis of Oral Cancer Employing Deep Learning Algorithm. *Journal of Cancer Research and Clinical Oncology*, **145**, 829-837. https://doi.org/10.1007/s00432-018-02834-7

[3] Wang, W., Liang, D. and Chen, Q. (2020) Medical Image Classification Using Deep Learning. *Deep Learning in Healthcare*, **171**, 33-51. https://doi.org/10.1007/978-3-030-32606-7_3

[4] Zou, W., Lu, H., Yan, K. and Ye, M. (2019) Breast Cancer Histopathological Image Classification Using Deep Learning. 2019 10*th International Conference on Information Technology in Medicine and Education* (*ITME*), Qingdao, 23-25 August 2019, 53-57. https://doi.org/10.1109/ITME.2019.00023

[5] Lu, X. and FiroozehAbolhasani Zadeh, Y.A. (2022) Deep Learning-Based Classification for Melanoma Detection Using XceptionNet. *Journal of Healthcare Engineering*, **2022**, Article ID: 2196096. https://doi.org/10.1155/2022/2196096

[6] Mahin, M., Tonmoy, S. and Islam, R. (2021) Classification of COVID-19 and Pneumonia Using Deep Transfer Learning. *Journal of Healthcare Engineering*, **2021**, Article ID: 3514821. https://doi.org/10.1155/2021/3514821

[7] Xue, Z., Yu, X. and Liu, B. (2021) HResNetAM: Hierarchical Residual Network with Attention Mechanism for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **14**, 3566-3580. https://doi.org/10.1109/JSTARS.2021.3065987

[8] Xu, L., Huang, J., Nitanda, A., *et al.* (2020) A Novel Global Spatial Attention Mechanism in Convolutional Neural Network for Medical Image Classification.

[9] Zou, Y., Zhang, J., Huang, S., *et al.* (2022) Breast Cancer Histopathological Image Classification Using Attention High-Order Deep Network. *International Journal of Imaging Systems and Technology*, **32**, 266-279. https://doi.org/10.1002/ima.22628

[10] Ning, D., Liu, G., Jiang, R., *et al.* (2019) Attention-Based Multi-Scale Transfer ResNet for Skull Fracture Image Classification. 4*th International Workshop on Pattern Recognition*, Vol. 11198, 63-67. https://doi.org/10.1117/12.2540498

[11] Wang, S., Li, R., Wang, X., *et al.* (2021) Multiscale Residual Network Based on Channel Spatial Attention Mechanism for Multilabel ECG Classification. *Journal of Healthcare Engineering*, **2021**, Article ID: 6630643. https://doi.org/10.1155/2021/6630643

[12] Du, W., Rao, N., Dong, C., *et al.* (2021) Automatic Classification of Esophageal Disease in Gastroscopic Images Using an Efficient Channel Attention Deep Dense Convolutional Neural Network. *Biomedical Optics Express*, **12**, 3066-3081.

https://doi.org/10.1364/BOE.420935

[13] Szegedy, C., Liu, W., Jia, Y., *et al.* (2015) Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[14] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[15] Ioffe, S. and Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning*, Vol. 37, 448-456.

[16] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[17] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[18] Woo, S., Park, J., Lee, J.Y., *et al.* (2018) CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[19] Ma, C., Wang, H. and Hoi, S.C.H. (2019) Multi-Label Thoracic Disease Image Classification with Cross-Attention Networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Vol. 11769, 730-738. https://doi.org/10.1007/978-3-030-32226-7_81

[20] Wang, F., Jiang, M., Qian, C., *et al.* (2017) Residual Attention Network for Image Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 3156-3164. https://doi.org/10.1109/CVPR.2017.683

[21] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 26 June-1 July 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[22] Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. (2017) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 4278-4284. https://doi.org/10.1609/aaai.v31i1.11231

[23] Yang, S., Lin, G., Jiang, Q. and Lin, W. (2020) A Dilated Inception Network for Visual Saliency Prediction. *IEEE Transactions on Multimedia*, **22**, 2163-2176. https://doi.org/10.1109/TMM.2019.2947352

[24] Zhang, X., Zou, Y. and Shi, W. (2017) Dilated Convolution Neural Network with LeakyReLU for Environmental Sound Classification. 2017 22*nd International Conference on Digital Signal Processing* (*DSP*), London, 23-25 August 2017, 1-5. https://doi.org/10.1109/ICDSP.2017.8096153

[25] Yan, C., Luo, Z., Lin, Z., *et al.* (2022) Shear Wave Elastography-Assisted Ultrasound Breast Image Analysis and Identification of Abnormal Data. *Journal of Healthcare Engineering*, **2022**, Article ID: 5499354. https://doi.org/10.1155/2022/5499354

[26] Kumar, M.D., Babaie, M., Zhu, S., *et al.* (2017) A Comparative Study of CNN, BoVW and LBP for Classification of Histopathological Images. 2017 *IEEE Symposium Series on Computational Intelligence* (*SSCI*), Honolulu, 27 November-1 December 2017, 1-7.

[27] Lin, T.Y., Goyal, P., Girshick, R., *et al.* (2017) Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.324

[28] Xiong, R., Yang, Y., He, D., *et al.* (2020) On Layer Normalization in the Transformer Architecture. *International Conference on Machine Learning*, 13-18 July 2020, 10524-10533.

[29] Spanhol, F.A., Oliveira, L.S., Petitjean, C., *et al.* (2016) Breast Cancer Histopathological Image Classification Using Convolutional Neural Networks. *International Joint Conference on Neural Networks* (*IJCNN*) *IEEE*, Vancouver, 24-29 July 2016, 2560-2567. https://doi.org/10.1109/IJCNN.2016.7727519

[30] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[31] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. 26*th Annual Conference on Neural Information Processing Systems* 2012, Lake Tahoe, 3-6 December 2012, 1097-1105.

[32] de Matos, J., de Souza Britto, A., de Oliveira, L.E.S., *et al.* (2019) Texture CNN for Histopathological Image Classification. 2019 *IEEE* 32*nd International Symposium on Computer-Based Medical Systems* (*CBMS*), Cordoba, 5-7 June 2019, 580-583. https://doi.org/10.1109/CBMS.2019.00120

[33] Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K. and Li, S. (2017) Breast Cancer Multi-Classification from Histopathological Images with Structured Deep Learning Model. *Scientific Reports*, **7**, Article No. 4172. https://doi.org/10.1038/s41598-017-04075-z

[34] Sharma, S., Mehra, R. and Kumar, S. (2021) Optimised CNN in Conjunction with Efficient Pooling Strategy for the Multi-Classification of Breast Cancer. *IET Image Processing*, **15**, 936-946. https://doi.org/10.1049/ipr2.12074

[35] Srikantamurthy, M.M., Rallabandi, V.P.S., Dudekula, D.B., *et al.* (2023) Classification of Benign and Malignant Subtypes of Breast Cancer Histopathology Imaging Using Hybrid CNN-LSTM Based Transfer Learning. *BMC Medical Imaging*, **23**, Article No. 19. https://doi.org/10.1186/s12880-023-00964-0

[36] Barzekar, H. and Yu, Z. (2022) C-Net: A Reliable Convolutional Neural Network for Biomedical Image Classification. *Expert Systems with Applications*, **187**, Article ID: 116003. https://doi.org/10.1016/j.eswa.2021.116003