Order-ViT: Order Learning Vision Transformer for Cancer Classification in Pathology Images

Ju Cheon Lee and Jin Tae Kwak
School of Electrical Engineering, Korea University, Seoul, Republic of Korea
dlwncjs0618@korea.ac.kr, jkwak@korea.ac.kr

Abstract

In computational pathology, cancer classification is one of the most widely studied tasks. There exist numerous tools for cancer classification, which are mainly built based upon convolutional neural networks or Transformers. These tools, by and large, formulate cancer classification as a categorical classification problem, which ignores the intrinsic relationship among cancer grades. Herein, we propose an order learning vision transformer for cancer classification that can not only learn the histopathological patterns of individual cancer grades but also utilize the ordering relationship among cancer grades. Built based upon vision transformer, the proposed method simultaneously conducts categorical classification per input sample and order classification for a pair of input and reference samples. Moreover, it introduces a voting scheme to identify less confident samples and to improve the accuracy of the decision on such samples. The proposed method is evaluated on two types of cancer datasets including colorectal and gastric cancers. Experimental results show that the proposed method outperforms other classification models and can facilitate improved cancer diagnosis in clinics.

1. Introduction

Cancer is one of the leading causes of death that affects health worldwide [3]. Rapid and accurate diagnosis of cancer is critical to initiating timely treatment and improving patient outcomes. The current gold standard for cancer diagnosis involves the microscopic examination of tissue samples stained with hematoxylin and eosin (H&E) by pathologists. However, these manual histopathological evaluations are time-consuming, low-throughput, and prone to inter- and intra-observer variability, and thus can limit the efficiency and accuracy of cancer diagnosis. Therefore, an automated, fast, and reliable methods are needed to improve the quality of cancer diagnosis and patient care.

Computational pathology, which combines artificial intelligence (AI), whole-slide imaging (WSI), and clinical informatics, is an emerging field of study that hold great potential for transforming and improving the practice of current pathology [2, 4, 8]. Many of these tools rely on deep learning algorithms, especially convolutional neural networks (CNNs). CNNs have been successfully applied to numerous applications including cancer diagnosis [43], tissue segmentation [16], and synthetic tissue image generation [38]. Recently, Vision Transformer (ViT) has received considerable attention for its excellent performance in computer vision tasks such as object detection [5], semantic segmentation [30], and scene understanding [40]. The application of ViT has been also extended to pathological image analysis such as histopathology image classification [10,34] metastasis detection [4], and survival analysis [23].

For both CNN-based and ViT-based models, cancer diagnosis has been primarily studied as a categorical classification problem. As a categorical classification problem, the role of CNNs or ViT is to identify the unique and distinct patterns of each cancer grade within a tissue image. In this perspective, Cancer grades are considered to be independent to each other. Although this approach has been successful, it, by and large, ignores the intrinsic relationships among cancer grades. It is generally accepted that the higher the cancer grade, the more aggressive it is. This indicates that there is an intrinsic ordering among cancer grades. [18], in fact, pointed this out and proposed a CNN-based model that conducts cancer grading as both categorical classification and regression problems, which handles one tissue image at a time. Moreover, the ordering relationship can be incorporated into cancer diagnosis via pairwise comparison between tissue samples. Suppose that there are borderline cases between two cancer grades such as high-grade cancer and low-grade cancer, potentially leading to mis-diagnosis in the clinics. For such cases, it will be easier or helpful to make a comparison against some representative cases of the two grades and tell whether is less or more aggressive than the representative cases. This approach of learning the ordering relationship is called order learning [20].

In this work, we propose an order learning ViT (Order-

ViT) for cancer grading in pathology images. Order-ViT conducts two classification tasks simultaneously: one for categorical classification and the other for order classification. For the categorical classification, it receives one tissue sample and predicts its class label, i.e., learning and exploiting specific histopathological patterns of cancer grades. As for order classification, it takes an additional reference tissue sample and predicts the ordering relationship between the two sample via pairwise comparison. Order-ViT contains a feature extractor, a categorical classifier, and an order classifier. We employ both CNN and ViT to build a feature extractor that extracts the high-dimensional feature representation of a tissue sample. Subsequently, the categorical classifier assigns a cancer grade per tissue sample and the order classifier receives the feature representations of a pair of tissue samples and predicts the ordering relationship between them. The categorical classifier conducts the primary classification. If and only if the result of the primary classification is uncertain, we invoke a voting scheme that conducts a number of ordinal classifications and voting to determine the final prediction.

Our contributions are summarized as follows:

- We propose an order learning ViT for cancer grading in pathology images, which learns both the histopathological patterns and the ordering relationship among cancer grades.
- We introduce a voting scheme that exploits the ordering relationship between tissue samples to further improve the accuracy and reliability of cancer grading.
- We evaluate the proposed method on two types of cancer datasets and compared it with CNN- and ViT-based models, outperforming other models regardless of the datasets.

2. Related Work

2.1. Computational Pathology for Cancer Diagnosis

A wide range of approaches and methods in computational pathology have been applied to cancer diagnosis. Early works focused on extracting and utilizing hand-crafted features that are associated with the presence and aggressiveness of cancer. The hand-crafted features include color intensity/histograms, morphological features, and texture features. Morphological features are to quantify the shape, distribution, and density of histological objects such as glands [29] and nuclei [14]. Texture features include grey-level co-occurrence matrix [27], wavelet transform [6], local binary pattern [25], and Gabor filters [24]. Later, deep learning methods, in particular CNNs, have shown to be effective in processing and analyzing pathology images. Various kinds of CNN architectures such as AlexNet [17], ResNet [9], EfficientNet [31], and MobileNet [11], which were developed for computer vision tasks, have been successfully applied to cancer diagnosis for colon cancer [26], gastric cancer [39], and breast cancer. In addition to these, a number of advanced models that are specific to pathology image analysis have been proposed. For example, [33] proposed a multi-scale CNN that extracts and utilizes binary patterns of feature maps across multiple scales for colorectal and prostate cancer classification. [19] adopted dual-domain attention mechanism to strengthen feature representations and to focus on important areas, achieving improved classification performance on colon and lung cancers. [42] introduced attention high-order model that simultaneously integrates the high-order statistics and attention mechanism modules into a residual network for breast cancer classification. Recently, Transformer has been adopted for several computer vision tasks as well as pathology image classification tasks, outperforming CNN-based models.

2.2. Vision Transformer (ViT)

Transformer was first introduced in the machine translation task in natural language processing. Vision Transformer (ViT), which is a variant of Transformer, was developed for the image classification task. ViT unfolds an input image as a series of small patches, i.e., tokens, so that Transformer layers can be utilized to incorporate the global context of the input image and to conduct the image classification task. This approach has been successfully adopted for various computer vision tasks, including image segmentation [15], image generation [37], and change detection of a pair of co-registered images [1]. It has been also widely applied to pathology image analysis. For example, [13] employed ViT for prostate cancer grading in WSIs by extracting and utilizing multiple image patches from WSIs. [41] conducted tissue segmentation in colorectal cancer images by adopting Swin Transformer for extracting the global context of tumour-related regions and a cascaded upsampler for detecting the tumor boundary. [35] proposed a hybrid architecture of Swin transformer and CNN that is equipped with contrastive learning to enhance feature extraction and to conduct five downstream tasks such as patch retrieval, patch classification, WSI classification, mitosis detection, and colorectal adenocarcinoma gland segmentation. [21] also exploited both CNN and ViT models to extract features from pathological images and then combine them to perform the classification of cervical cancer images.

2.3. Order Learning

Order learning is a learning paradigm that aims at learning the ordered relationship among class labels, which was first introduced in age estimating [20]. Pairwise comparison is one of the most common ways to learn the order or rank relationship among samples. For instance, [20] employed a

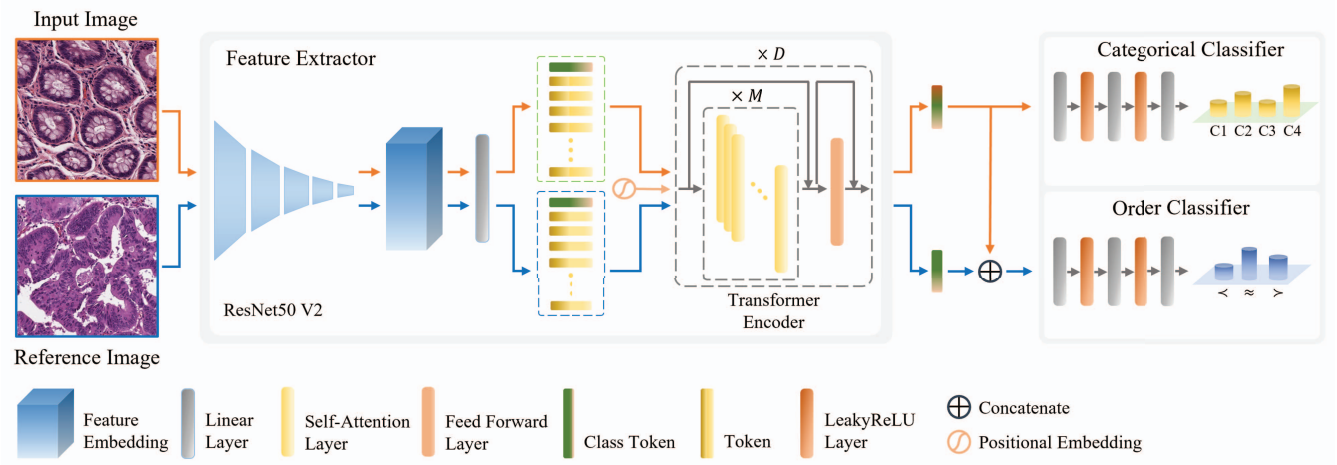

Figure 1. The overiview of the proposed Order-ViT during training. For the optimization of Order-ViT, we utilize a feature extractor, a categorical classifier, and an order classifier. The categorical classifier conducts classification using the input image alone. The order classifier compares the input image against the reference image.

pariwise comparator to compare two samples and to determine their ordering relationship. [36] developed RankNet that learns a ranking function to predict the ranking posterior of two samples. These methods generally utilize the absolute rank of a sample. In order to learn the relative rank that quantifies the degree of difference, [28] proposed a regression-based order learning mechanism for age estimation. Although it has been successfully applied to age estimation, it has not been utilized for pathology image analysis. To the best of our knowledge, this is the first attempt to incorporate order learning into cancer diagnosis in pathology images. Ordinal classification or regression approaches, which aim at directly predicting the class label of a sample on an ordinal scale, have been adopted for pathology image analysis; for example, [18] conducted both categorical classification and regression for prostate and colorectal cancer grading. However, the ordering relationships among samples and class labels have not been explicitly utilized for cancer diagnosis in pathology images.

3. Methods

3.1. Problem Formulation

Suppose that we are given N pathology images and their ground truth labels $\{x_i, y_i\}_{i=1}^{N}$ where $x_i \in R^3$ is the i-th pathology image, $y_i \in \mathcal{C} = \{c_1, c_2, \ldots, c_{N_c}\}$ is the ground truth class label for x_i. $c_1, c_2, \ldots, c_{N_c}$ represent cancer grades, of which each owns a unique histopathology pattern, and are also defined on an ordinal scale $c_1 = 1 < c_2 = 2 < \ldots < c_{N_c} = N_c$. Given $\{x_i, y_i\}_{i=1}^{N}$, Order-ViT maps the images into a high-dimensional embedding space in which we conduct two classification tasks, i.e., categorical classification and order classification. The objective of our study is to learn Order-ViT that can leverage the ordering relationships among pathology images and conduct cancer grading in an accurate and robust manner, which can be formulated as follows:

$$min \sum_{i=1}^{N} \mathcal{L}(\xi(x_i), y_i; \theta) \tag{1}$$

where $\mathcal{L}$ is the loss function and θ is a set of learnable parameters of Order-ViT ξ.

3.2. Order-ViT

Order-ViT ξ contains four major computational blocks: 1) a feature extractor ξ^{feat}, a categorical classifier ξ^{cat}, an order classifier ξ^{ord}, and a voting block ξ^{vot}. Given an input image x, the feature vector ξ^{feat} extracts a feature vector f, which is subsequently fed into the categorical classifier ξ^{cat} for conducting categorical classification, i.e., directly predicting the class label for the x. For order learning, a reference image x^{ref} goes through ξ^{feat}, generating a reference feature vector f^{ref}. f and f^{ref} are concatenated together and fed into the order classifier ξ^{ord} to predict the order relationship between x and x^{ref}. The voting block ξ^{vot} is invoked at inference only to identify hard/uncertain images and to correct and stabilize the prediction.

3.2.1 Feature Extractor

The feature extractor ξ^{feat} is built based upon a vision transformer (ViT). Suppose that we are given an input image x, ξ^{feat} adopts Resnet50V2 [9] to extract the initial feature vector $e \in \mathcal{R}^{W \times H \times N_e} (W = H = 24, N_e = 1024)$ and generates a set of token embeddings via a linear layer $\{s_i | i = 1, 2, \ldots, N_s\}$, $s_i \in \mathcal{R}^{N_f}$ where N_s is the length of the token embeddings ($N_s = 576$) and N_f is the size of each token embedding ($N_f = 768$). For classification,

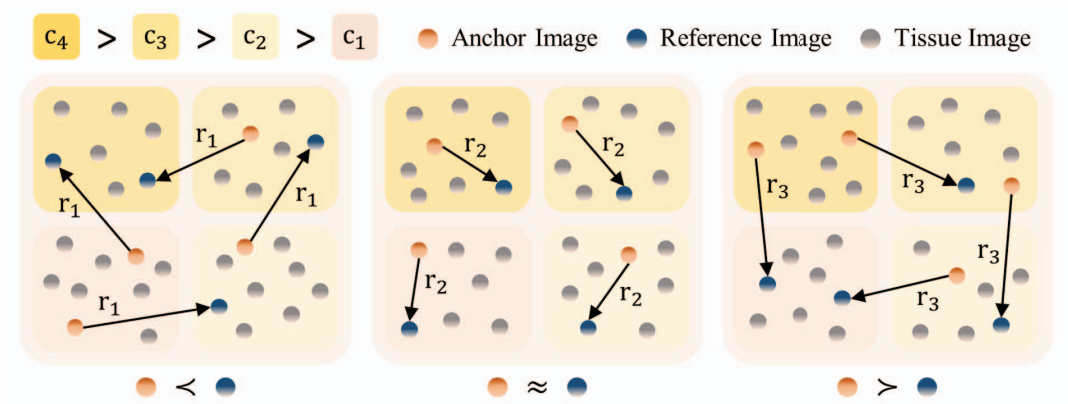

Figure 2. Generation of order class labels. The class label of an anchor image is compared against that of a reference image to determine whether it is higher than $\succ$, equal to $\approx$, or lower than $\prec$ the reference image, corresponding to the order class label r_1, r_2, and r_3, respectively.

we incorporate a class token $s^{CLS} \in \mathcal{R}^{N_f}$ into the token embeddings, producing $H_o = [s_0, s_1, s_2, .., s_{N_s}], s_0 = s^{CLS}$. We also utilize positional embedding to add location information for each token embedding in a way that $\sin(i/10000^{2j+1/N_f})$ and $\cos(i/10000^{2j/N_f})$ are added to every $2j$-th and $2j+1$-th dimensions of $s_i, i = 0, 1, ..., N_s$. The resultant token embeddings are fed to the Transformer encoder, which is built based upon multi-head self-attention ($MHSA$) layers, feed-forward neural network ($FFNN(\cdot)$) layers, and a layer normalization (LN) layer, to extract the output feature f as follows:

$$h'_l = MHSA(LN(h_{l-1})) + h_{l-1}, l = 1, ..., N_l \quad (2)$$

$$h_l = FFNN(LN(h_l)) + h'_l, l = 1, ..., N_l \quad (3)$$

$$f = LN(h^0_{N_l}) \quad (4)$$

where B_l is the number of Transformer encoders ($N_l = 12$) and $h^0_{N_l}$ is the output embedding corresponding to s^{CLS} at the N_l-th Transformer encoder. $MHSA$ is composed of multiple, parallel self-attentions (SAs) as follows:

$$MHSA(h) = Linear([SA_1(h), ..., SA_{N_m}(h)]) \quad (5)$$

$$SA_i(h) = softmax(qk^T/\sqrt{D})v, i = 1, ..., N_m \quad (6)$$

where $q, k, v = Linear(h) \in \mathcal{R}^{d_{N_m}}$ are query, key, value vectors, respectively, $Linear$ is a linear layer, N_m is the number of SAs, and $D = N_f/N_m$. $FFNN$ contains $Linear - ReLU - Linear - ReLU - Linear$.

3.2.2 Categorical Classifier

The categorical classifier ξ^{cat} contains $Linear - LeakyReLU - Linear - LeakyReLU - Linear$. Provided with a feature vector f, ξ^{cat} produces the posterior probabilities for the class labels $\hat{p} = \{\hat{p}^c | c = 1, ..., N_c\}$ where $\hat{p}^c$ denotes the probability that the input image belongs to the c-th class label.

3.2.3 Order Classifier

Following [20], the order relationship between two samples x_i and x_j can be determined by using their class labels y_i and y_j as follows:

$$\begin{aligned} x_i \prec x_j \quad &\text{if} \quad y_i - y_j < 0 \\ x_i \approx x_j \quad &\text{if} \quad y_i - y_j = 0 \\ x_i \succ x_j \quad &\text{if} \quad y_i - y_j > 0 \end{aligned} \quad (7)$$

where $x_i \succ x_j$, $x_i \approx x_j$, and $x_i \prec x_j$ indicate that x_i is greater than, equal to, or less than x_j with respect to the class labels that are defined on an ordinal scale. In the context of cancer grading, these are interpreted as the grade of x_i is higher than, equal to, or lower than that of x_j. Given the two feature vectors from x_i and x_j, the objective of the order classifier ξ^{ord} is to predict the order relationship ($\succ, \approx, \prec$) between x_i and x_j that is defined as follows:

$$z_{i,j} = \begin{cases} r_1, & \text{if } y_i < y_j \\ r_2, & \text{if } y_i = y_j \\ r_3, & \text{if } y_i > y_j \end{cases} \quad (8)$$

where $r_1, r_2,$ and $r_3 \in \mathcal{N}$ are the ground truth order class labels that correspond to $\succ, \approx,$ and $\prec$, respectively.

During training, each image in a batch is designated as an anchor image x^a and ξ^{feat} is used to generate an anchor feature vector f^a. Then, a reference feature vector f^{ref} is selected from a feature memory bank $\mathcal{M}$ and is concatenated with f^a, producing the order feature vector $f^{ord} = f^a \oplus f^{ref}$. xi^{ord} receives f^{ord} and produces the posterior probabilities for the order class labels $\hat{o} = \{\hat{o}^i | i = 1, ..., N_o\}$ where $\hat{o}^i$ denotes the probability that the input image belongs to the i-th order class label.

The architecture of ξ^{ord} is identical to that of ξ^{cat} except the number of neurons in the last linear layer.

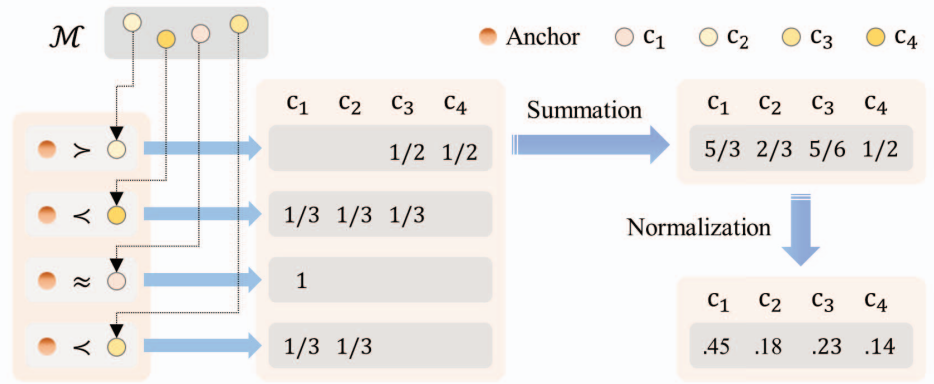

Figure 3. Voting at inference. The feature vector of an anchor image is compared to a number of reference feature vectors in the feature memory bank $\mathcal{M}$ to conduct voting. The voting results are summarized to produce posterior probabilities.

3.3. Two-stage Inference

During inference, we make the final decision or prediction in two-stages. In the first stage, we make the prediction per input image x^a using the categorical classifier ξ^{cat}. Specifically, we extract the feature vector f^a for x^a by using the feature extractor ξ^{feat}. Then, f^a is fed into ξ^{cat} to conduct categorical classification. Using the output of ξ^{cat}, we also calculate a confidence score u of the input image. If the prediction is certain, i.e., the confidence score $u > \tau$, then the prediction by ξ^{cat} becomes the final prediction. τ is a threshold value. Otherwise, we invoke a voting block ξ^{vot} in the second stage. ξ^{vot} utilizes the order classifier ξ^{ord} using a feature memory bank $\mathcal{M} = \{\mathcal{M}^c | c = c_1, c_2, ..., c_{N_c}\}$ where $\mathcal{M}^c$ is a set of feature vectors of class c. For each feature vector $f^m \in \mathcal{M}^c$, we create the order feature vector $f^{ord} = f^a \oplus f^m$ and fed it into the order classifier ξ^{ord}, producing the order relationship $(\succ, \approx, \prec)$ between them. We accumulate the order classification results for $\mathcal{M}$ and determine the final prediction.

3.3.1 Confidence Score

We compute the confidence score u of the input image by obtaining the posterior probabilities for the class labels $\hat{p} = \{\hat{p}^i | i = 1, ..., N_c\}$, given by ξ^{cat}, identifying the highest probability $\hat{p}_{first}$ and the second highest probability $\hat{p}_{second}$, and calculating the difference between them, which is designated as the confidence score $u = \hat{p}_{first} - \hat{p}_{second}$. The larger the difference between the two probabilities is, the more confident the decision is.

To determine whether we invoke the voting scheme ξ^{vot} or not, we calculate the threshold value τ using the images in the validation set. From the validation set, we collect the images that are mis-classified by ξ^{cat}, which forms $\{x_i | i = 1, ..., N_m\}$ where N_m is the number of mis-classified images in the validation set. For these images, we compute the confidence scores $\{u_i | i = 1, ..., N_m\}$ as described above and sort the images in decreasing order by the confidence scores. Using each confidence score as a threshold value τ^*, we obtain the images that are smaller than τ^*, i.e., the images that the model is less confident. The number of such images is designated as N_w. Then, we select the threshold value τ as the largest τ^* that produces $\frac{N_w}{N_m} \leq 0.1$.

3.3.2 Feature Memory Bank

The feature memory bank $\mathcal{M} = \{\mathcal{M}^c | c = c_1, c_2, ..., c_{N_c}\}$ consists of a number of feature vectors per class label. To construct $\mathcal{M}$, we select N_r reference images per class label from the training and validation sets that are representative of each class. ξ^{cat} classifies each image in the training and validation sets. If the result is correct and its highest probability is 0.9 or higher, then it is designated as a candidate sample. Once we obtain all the candidate samples, we randomly select N_r images per class label, extract their feature vectors, and construct $\mathcal{M}$. We set N_r to 10.

3.3.3 Voting

Given f^a of an uncertain image, i.e., anchor image x^a, and $\mathcal{M}$, ξ^{vot} uses ξ^{ord} to conduct a pair-wise comparison between x^a and every reference image x^{ref} that is represented by the feature vector in $\mathcal{M}$. The results of ξ^{ord}, i.e., $\succ, \approx$, and $\prec$ indicate that x^a is ranked higher than, equal to, and lower than x^{ref}, respectively. Suppose that the ground truth class label of x^{ref} is c. We vote for the prediction in three different manners: 1) if ξ^{ord} produces $\succ$, we add $1/(N_c - c)$ for all labels higher than c, 2) if the result is $\prec$, we add $1/(N_c - c)$ for all labels lower than c, and 3) if the result is $\approx$, we add 1.0 to c. By normalizing the votes, we construct the $q = \{q^c | c = 1, ..., N_c\}$ where q^c represents the probability that x^a belongs to the c-th class label.

For x^a, two posterior probabilities are available, including $\hat{p}$ by ξ^{cat} and q by ξ^{cat}. To make the final prediction, we conduct a weighted sum of the two posterior probabili-

ties and take the class label with the highest probability. The weights (w_1 and w_2) are adaptively computed as follows:

$$w_1 = 1 - \frac{H(\hat{p})}{H(q) + H(\hat{p})}, w_2 = 1 - \frac{H(q)}{H(q) + H(\hat{p})} \quad (9)$$

where H is an entropy. The less uncertain the probability distribution is, the higher weight it gets. The final probabilities q_{final} are calculated as $q_{final} = \hat{p} \times w_1 + q \times w_2$.

3.4. Loss Function

The objective function for Order-ViT is given by

$$L_{total} = L_{cat} + \lambda L_{ord} \quad (10)$$

where L_{cat} and L_{ord} denotes the loss function for category classification and order classification, respectively, and λ is a hyper-parameter to balance the effect of the two loss functions ($\lambda = 0.4$). Both category classification and order classification adopt cross-entropy loss.

4. Experiment

Table 1. Details of colorectal and gastric tissue datasets.

Tissue Type	Class	Training	Validation	TestI	TestII
Colorectal Tissue	BN	773	374	453	27,896
	WD	1,866	264	192	8,394
	MD	2,997	370	738	61,985
	PD	1,391	234	205	11,896
Gastric Tissue	BN	20,883	8,398	7,955	-
	WD	14,251	2,239	1,795	-
	MD	20,815	2,370	2,458	-
	PD	27,689	2,374	3,579	-

4.1. Dataset

We employ two kinds of pathology image datasets, including colorectal tissue dataset and gastric tissue dataset. The details of the datasets are described in Table 1.

4.2. Colorectal Tissue Dataset

Two sets of colorectal tissue datasets were obtained from [18]. The first data set was obtained from 6 colorectal tissue microarrays (TMAs) (0.2465 µm × 0.2465 µm) and 3 WSIs (0.2518 µm × 0.2518 µm), at 40x magnification using an Aperio digital slide scanner (Leica Biosystems). This contains a training set, a validation set, and a test set (C_{TestI}). The second dataset (C_{TestII}) was obtained from 45 WSIs (0.2253 µm × 0.2253 µm) at 40× magnification using a NanoZoomer digital slide scanner (Hamamatsu Photonics K.K.). For both datasets, tissue image patches (1024 × 1024 pixels) are annotated with benign (BN), well-differentiated (WD), moderately- differentiated (MD), and poorly-differentiated (PD).

4.3. Gastric Tissue Dataset

98 anonymized, gastric WSIs, digitized at 40× magnification using an Aperio digital slide scanner (Leica Biosystems) (0.2635 µm × 0.2635 µm), were obtained from a local hospital. Upon pathologic review, 114,842 image patches (1024 × 1024 pixels) were obtained and categorized into BN, tubular well-differentiated (TW) tumor, tubular moderately-differentiated (TM), and tubular poorly-differentiated (TP) tumor. These constitute a training set, a validation set, and a test set (G_{Test}).

4.4. Comparative Experiments

We compare Order-ViT with a number of CNN- and Transformer-based models. Three types of CNN-based models are employed: 1) three plain CNNs: ResNet50 [9], DenseNet121 [12], and EfficientNetB0 [31], 2) a multi-scale CNN: MSBP-Net [33] MSBP-Net, and 3) two multi-task CNNs [18]: $\mathcal{M}_{MSE-CE_o}$ and $\mathcal{M}_{MAE-CE_o}$. Three Transformer-based models are utilized: 1) ViT [7], 2) Swin-transformer [22], and 3) DeiT III [32].

4.5. Evaluation Metrics

Three evaluation metrics are adopted: 1) accuracy (Acc): measures the ratio of correctly classified instances to the total number of instances, 2) macro F1-score (F1): a harmonic mean of the average precision and average recall, and 3) quadratic weighted kappa ((k_w): quantifies the degree of agreement between the prediction and ground truth labels.

4.6. Implementation Details

We implemented all models using the Pytorch platform with two RTX 3090 GPUs. All the models adopted pre-trained weights from ImageNet and were trained with 50 epochs using a batch size of 16, Adam optimizer (β_1 : $0.9, \beta_2$: $0.999, \epsilon$: $1.0e^{-}8$), and Cosine annealing warm restarts scheduler (learning rate: $1.0e^{-}4$). All the tissue images were resized to 384 x 384 pixels and 512 x 512 pixels Transformer-based and CNN-based models, respectively. Several data augmentation techniques, including Affine transformation, Gaussian blur, average blur, median blur, Gaussian noise, dropout, linear contrast, horizontally and vertically flipping, and random saturation, were applied using Aleju library.

5. Result

5.1. Colorectal Cancer Classification

Table 2 shows the results of colorectal cancer classification (C_{TestI} and C_{TestII}) by Order-ViT and other competing models. On C_{TestI}, Order-ViT achieved the best Acc of 87.66% and k_w of 0.942 and the second best F1 of 0.834, outperforming all competitors except F1 by

Table 2. Result of colorectal and gastric cancer classification.

Model	C_{TestI}			C_{TestII}			G_{Test}		
	Acc(%)	F1	k_w	Acc(%)	F1	k_w	Acc(%)	F1	k_w
ResNet	86.65	0.825	0.935	68.24	0.684	0.818	84.04	0.775	0.926
DenseNet	85.77	0.820	0.930	70.03	0.671	0.843	83.44	0.772	0.928
EfficientNet	86.71	0.821	0.926	64.60	0.627	0.794	82.81	0.766	0.919
MSBP-Net	86.21	0.824	0.933	74.67	0.708	0.860	84.53	0.770	0.928
M_{MSE-CE_o}	87.28	**0.838**	0.940	75.95	0.710	0.846	84.34	0.771	0.925
M_{MAE-CE_o}	86.65	0.826	0.937	75.44	0.702	0.861	84.10	0.768	0.922
ViT	86.27	0.829	0.931	77.54	0.712	0.874	84.06	0.772	**0.931**
Swin	85.26	0.820	0.931	77.10	0.721	0.868	83.71	0.759	0.919
DeiT III	76.76	0.673	0.794	48.42	0.396	0.271	77.05	0.656	0.847
Order-ViT (Ours)	**87.66**	0.834	**0.942**	**83.21**	**0.740**	**0.899**	**84.89**	**0.783**	0.930

$\mathcal{M}_{MSE-CE_o}$. Among the competitors, two multitask CNN models ($\mathcal{M}_{MSE-CE_o}$ and $\mathcal{M}_{MAE-CE_o}$) were superior to others. Among Transformer-based models, ViT outperformed two other Transformer-based models. DeiT III was inferior to all models regardless of the evaluation metrics. On C_{TestII}, Order-ViT outperformed all competitors regardless of the evaluation metrics. Among other competitors, two multitask CNN models ($\mathcal{M}_{MSE-CE_o}$ and $\mathcal{M}_{MAE-CE_o}$) and ViT were still the best performing models. We note that C_{TestII} was obtained under different acquisition settings from the training set. In a head-to-head comparison between the two datasets, there was a consistent performance drop for all models under consideration. However, Order-ViT showed the least performance drop in comparison to other models. For instance, 4.45% Acc, 0.094 F1, and 0.043 k_w were dropped by Order-ViT. 8.16 $\sim$ 28.34% Acc, 0.099 $\sim$ 0.277 F1, and 0.057 $\sim$ 0.523 k_w were dropped by other models.

5.2. Gastric Cancer Classification

Table 2 describes the performance of Order-ViT and other competitors on gastric cancer classification (G_{Test}). Order-ViT achieved the best Acc of 84.89% and F1 of 0.783 and second best k_w of 0.930. Among other models, mixed results were observed; MSBP-Net, ResNet, and ViT obtained the best Acc of 84.04%, F1 of 0.775, and k_w of 0.931, respectively. DeiT III was still the worst model in all evaluation metrics.

5.3. Ablation Study

We conducted a series of ablation experiments on Order-ViT to investigate the effect and role of its computational components and design choices.

Effect of computational blocks. First, we assessed the effect of the computational blocks, including the order classifier ξ^{ord} and/or the voting block ξ^{vot}, on Order-ViT. Table 3 demonstrates the results of colorectal and gastric cancer classification without ξ^{ord} and/or ξ^{vot}. Order-ViT without ξ^{vot} experienced a consistent performance drop across different datasets; 0.25% Acc, 0.005 F1, and 0.001 k_w on C_{TestI}; 0.44% Acc, 0.003 F1, and 0.002 k_w on C_{TestII}; 0.31% Acc, 0.007 F1, and 0.001 k_w on G_{Test}. Eliminating both ξ^{ord} and ξ^{vot}, the performance, in general, further decreased except k_w on G_{Test}; 1.39% Acc, 0.005 F1, and 0.011 k_w on C_{TestI}, 5.67% Acc, 0.028 F1, and 0.025 k_w on C_{TestII}, and 0.83% Acc and 0.011 F1 on G_{Test}. Moreover, we assessed the effect of the weighted sum of the two posterior probabilities from the categorical and order classifications (Σ^w). Replacing Σ^w by addition, there is a slight performance drop regardless of the datasets and evaluation metrics. These results demonstrate that the effectiveness of the computational blocks in Order-ViT.

Effect of the number of reference images. Second, we investigated the effect of the number of reference images N_r in the voting block ξ^{vot}. We set N_r to 10, 30, 50, 70, 90, 100, and 200. As N_r changes, the classification performance was more or less the same across different datasets. This suggests that Order-ViT is not sensitive to the number of references images for voting. By setting N_r to 10, we can not only improve the accuracy and robustness of cancer classification, but also reduce the computational complexity during the inference.

5.4. Computational Complexity

Table 5 depicts the computational complexity of Order-ViT and other competing models. Including Order-ViT, Transformer-based models contain 2- to 18-fold more parameters than CNN-based models. Similar observations were made for the number of floating-point operations (FLOPs). The training time of the models varied regardless of the type of the models; however, CNN-based models, in general, demonstrated the shorter inference time than Transformer-based models. Due to the hybrid architecture of the feature extractor and voting mechanism, Order-ViT showed the most FLOPs and the longest inference time among the models under considerations.

Table 3. Ablation study on Order-ViT for colorectal and gastric cancer classification.

ξ^{cat}	ξ^{ord}	ξ^{vot}	Σ^w	C_{TestI}			C_{TestII}			G_{Test}		
				Acc(%)	F1	k_w	Acc(%)	F1	k_w	Acc(%)	F1	k_w
✓				86.27	0.829	0.931	77.54	0.712	0.874	84.06	0.772	**0.931**
✓	✓			87.41	0.829	0.941	82.77	0.737	0.897	84.58	0.776	0.929
✓	✓	✓		87.47	0.831	0.941	82.98	0.739	0.898	84.61	0.779	0.929
✓	✓	✓	✓	**87.66**	**0.834**	**0.942**	**83.21**	**0.740**	**0.899**	**84.89**	**0.783**	0.930

Table 4. Effect of N_r on colorectal and gastric cancer classification.

N_r	C_{TestI}			C_{TestII}			G_{Test}		
	Acc(%)	F1	k_w	Acc(%)	F1	k_w	Acc(%)	F1	k_w
10	**87.66**	**0.834**	**0.942**	**83.21**	**0.740**	**0.899**	**84.89**	**0.783**	**0.930**
30	87.47	0.831	0.940	83.20	0.740	0.899	84.87	0.782	0.929
50	87.47	0.831	0.940	83.20	0.740	0.899	84.85	0.782	0.929
70	87.59	0.832	0.941	83.20	0.740	0.899	84.88	0.782	0.929
90	87.59	0.832	0.941	83.20	0.740	0.899	84.85	0.782	0.929
100	87.53	0.832	0.941	83.19	0.739	0.899	84.88	0.782	0.929
200	87.53	0.832	0.941	83.19	0.739	0.899	84.88	0.782	0.929

Table 5. Computational complexity of Order-ViT and other competing models.

Model	FLOPs (M)	Params (M)	Training (ms/image)	Inference (ms/image)
ResNet	21587.17	24.69	18.13	3.09
DenseNet	15954.34	7.53	5.71	4.42
EfficientNet	152.22	4.81	3.31	3.12
MSBP-Net	44285.04	29.50	47.37	9.21
M_{MSE-CE_o}	141.14	4.84	27.44	4.28
M_{MAE-CE_o}	141.14	4.84	27.77	4.19
ViT	49505.12	86.37	27.01	7.84
Swin	44753.49	87.33	45.51	8.45
DeiT III	49608.92	86.17	25.84	6.27
Order-ViT	99046.44	87.29	27.95	9.76

5.5. Conclusions

In this work, we propose an efficient and effective cancer classification method that can leverage the individual histological patterns of pathology images by categorical classification and the relationship among different pathology images by order learning. The experimental results demonstrate that the proposed method can aid in improving the accuracy and robustness of cancer classification in pathology images. The future study will entail further development of order learning approaches for pathology image analysis and application of the proposed method to other types of organs and diseases.

Acknowledgements

This work was supported by the grant of the National Research Foundation of Korea (NRF) (No. 2021R1A2C2014557) and Institute of Information & communications Technology Planning & evaluation (IITP) (No. RS-2022-00167143), funded by the Korea government (MSIT).

References

[1] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 2

[2] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 1

[3] Freddie Bray, Mathieu Laversanne, Elisabete Weiderpass, and Isabelle Soerjomataram. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, 127(16):3029–3030, 2021. 1

[4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 1

[6] Liu Chun-Lin. A tutorial of the wavelet transform. *NTUEE, Taiwan*, 21:22, 2010. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[8] Babak Ehteshami Bejnordi, Maeve Mullooly, Ruth M Pfeiffer, Shaoqi Fan, Pamela M Vacek, Donald L Weaver, Sally Herschorn, Louise A Brinton, Bram van Ginneken, Nico Karssemeijer, et al. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Modern Pathology*, 31(10):1502–1512, 2018. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 6

[10] Zhu He, Mingwei Lin, Zeshui Xu, Zhiqiang Yao, Hong Chen, Adi Alhudhaif, and Fayadh Alenezi. Deconv-transformer (dect): A histopathological image classification model for breast cancer based on color deconvolution and transformer architecture. *Information Sciences*, 608:1093–1112, 2022. 1

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[13] Kobiljon Ikromjanov, Subrata Bhattacharjee, Yeong-Byn Hwang, Rashadul Islam Sumon, Hee-Cheol Kim, and Heung-Kook Choi. Whole slide image analysis and detection of prostate cancer using vision transformers. In *2022 international conference on artificial intelligence in information and communication (ICAIIC)*, pages 399–402. IEEE, 2022. 2

[14] Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2013. 2

[15] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2

[16] Sara Hosseinzadeh Kassani, Peyman Hosseinzadeh Kassani, Michal J Wesolowski, Kevin A Schneider, and Ralph Deters. Deep transfer learning based model for colorectal cancer histopathology segmentation: A comparative study of deep pre-trained models. *International Journal of Medical Informatics*, 159:104669, 2022. 1

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[18] Trinh Thi Le Vuong, Kyungeun Kim, Boram Song, and Jin Tae Kwak. Joint categorical and ordinal learning for cancer grading in pathology images. *Medical image analysis*, 73:102206, 2021. 1, 3, 6

[19] Han Li, Peishu Wu, Zidong Wang, Jingfeng Mao, Fuad E Alsaadi, and Nianyin Zeng. A generalized framework of feature learning enhanced convolutional neural network for pathology-image-oriented cancer diagnosis. *Computers in biology and medicine*, 151:106265, 2022. 2

[20] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *International Conference on Learning Representations*, 2020. 1, 2, 4

[21] Wanli Liu, Chen Li, Ning Xu, Tao Jiang, Md Mamunur Rahaman, Hongzan Sun, Xiangchen Wu, Weiming Hu, Haoyuan Chen, Changhao Sun, et al. Cvm-cervix: A hybrid cervical pap-smear image classification framework using cnn, visual transformer and multilayer perceptron. *Pattern Recognition*, 130:108829, 2022. 2

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[23] Zhilong Lv, Yuexiao Lin, Rui Yan, Zhenghe Yang, Ying Wang, and Fa Zhang. Pg-tfnet: transformer-based fusion network integrating pathological images and genomic data for cancer survival analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 491–496. IEEE, 2021. 1

[24] Javier R Movellan. Tutorial on gabor filters. *Open source document*, 40:1–23, 2002. 2

[25] Matti Pietikäinen. Local binary patterns. *Scholarpedia*, 5(3):9775, 2010. 2

[26] Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, and Pinkie Anggia. Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. *Procedia Computer Science*, 179:423–431, 2021. 2

[27] Bino Sebastian V, A Unnikrishnan, and Kannan Balakrishnan. Gray level co-occurrence matrices: generalisation and some new features. *arXiv preprint arXiv:1205.4831*, 2012. 2

[28] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18760–18769, 2022. 3

[29] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017. 2

[30] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 1

[31] X TanM and V LeQ. Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning. New York: IEEE*, volume 97, pages 6105–6114, 2019. 2, 6

[32] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 516–533. Springer, 2022. 6

[33] Trinh TL Vuong, Boram Song, Kyungeun Kim, Yong M Cho, and Jin T Kwak. Multi-scale binary pattern encoding network for cancer classification in pathology images. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1152–1163, 2021. 2, 6

[34] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021. 1

[35] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 2

[36] Xiaoshan Yang, Tianzhu Zhang, Changsheng Xu, Shuicheng Yan, M Shamim Hossain, and Ahmed Ghoneim. Deep relative attributes. *IEEE Transactions on Multimedia*, 18(9):1832–1842, 2016. 3

[37] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11304–11314, 2022. 2

[38] Qianqian Zhang, Haifeng Wang, Hongya Lu, Daehan Won, and Sang Won Yoon. Medical image synthesis with generative adversarial networks for tissue recognition. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 199–207. IEEE, 2018. 1

[39] Yuxue Zhao, Bo Hu, Ying Wang, Xiaomeng Yin, Yuanyuan Jiang, and Xiuli Zhu. Identification of gastric cancer with convolutional neural networks: a systematic review. *Multimedia Tools and Applications*, 81(8):11717–11736, 2022. 2

[40] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

[41] Usama Zidan, Mohamed Medhat Gaber, and Mohammed M Abdelsamea. Swincup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer. *Expert Systems with Applications*, 216:119452, 2023. 2

[42] Ying Zou, Jianxin Zhang, Shan Huang, and Bin Liu. Breast cancer histopathological image classification using attention high-order deep network. *International Journal of Imaging Systems and Technology*, 32(1):266–279, 2022. 2

[43] Juan Zuluaga-Gomez, Zeina Al Masry, Khaled Benaggoune, Safa Meraghni, and Nourredine Zerhouni. A cnn-based methodology for breast cancer diagnosis using thermal images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 9(2):131–145, 2021. 1