Uncertainty Ordinal Multi-Instance Learning for Breast Cancer Diagnosis

Xinzheng Xu 1,*, Qiaoyu Guo 1, Zhongnian Li 1 and Dechun Li 2

1 School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
2 Xuzhou Central Hospital, Xuzhou 221116, China
* Correspondence: xxzheng@cumt.edu.cn

Abstract: Ordinal multi-instance learning (OMIL) deals with the weak supervision scenario wherein instances in each training bag are not only multi-class but also have rank order relationships between classes, such as breast cancer, which has become one of the most frequent diseases in women. Most of the existing work has generally been to classify the region of interest (mass or microcalcification) on the mammogram as either benign or malignant, while ignoring the normal mammogram classification. Early screening for breast disease is particularly important for further diagnosis. Since early benign lesion areas on a mammogram are very similar to normal tissue, three classifications of mammograms for the improved screening of early benign lesions are necessary. In OMIL, an expert will only label the set of instances (bag), instead of labeling every instance. When labeling efforts are focused on the class of bags, ordinal classes of the instance inside the bag are not labeled. However, recent work on ordinal multi-instance has used the traditional support vector machine to solve the multi-classification problem without utilizing the ordinal information regarding the instances in the bag. In this paper, we propose a method that explicitly models the ordinal class information for bags and instances in bags. Specifically, we specify a key instance from the bag as a positive instance of bags, and design ordinal minimum uncertainty loss to iteratively optimize the selected key instances from the bags. The extensive experimental results clearly prove the effectiveness of the proposed ordinal instance-learning approach, which achieves 52.021% accuracy, 61.471% sensitivity, 47.206% specificity, 57.895% precision, and an 59.629% F1 score on a DDSM dataset.

Keywords: ordinal classification; multi-instance learning; weak supervision; breast cancer; key instance; uncertainty select

1. Introduction

Breast cancer is one of the most fatal diseases among women. The study of the benign and malignant classifications of breast cancer has been quite extensive. Elmoufidi [1] proposed a framework that uses a modified K-means algorithm to segment the ROI and extract textural features from the ROI for classification. Fahssi et al. [2] presented a novel CAD system for mammography diagnosis. The ROIs are detected by dividing the mammogram into regions and MIL algorithms are applied to identify malignant regions to label the whole mammogram. Most of the existing work has solved the binary classification problem of breast cancer, but there are more normal breast images than abnormal ones. Although the classification of benign and malignant regions of interest in abnormal breast images can achieve good performance, when the classification category becomes normal, benign, or malignant, the classification performance tends to decline. Lamard et al. [3] carried out experiments on binary classification and three-class classification. The experimental results show that the accuracy of the three-class classification is 30% lower than that of the two-class classification, which undoubtedly shows that the three-class classification of breast cancer is more challenging than the two-class classification. Moreover, we found...
that the classification of normal and benign regions is much more difficult than benign and malignant. The effective screening of early benign lesions is of great importance for further diagnosis and treatment. Since there is a certain order of lesions in the normal, benign, and malignant categories, we transform the general multi-instance learning problem into an orderly multi-instance-learning technique.

Multi-instance learning (MIL) [4–6] is a popular learning framework in weakly supervised learning [7–10]. Different from supervised learning, the data sample for MIL is a bag that consists of a set of instances. In the MIL setting, bags are labeled, while instances in bags are unlabeled. The researchers have studied various paradigms for MIL, including, but not limited to, multiple instance classification (MIC) [11], multiple instance regression, multi-instance clustering, imbalanced multi-instance learning, and multi-instance multi-labeling.

Here, we consider another typical scenario of MIL-ordinal multiple instance learning [12] (OMIL): the classes of bags are not only ordered, but the instances in the bag are also in a certain rank order. Moreover, the highest level of instance label cannot exceed the label of the bag. For example, in the medical diagnosis of breast cancer, the class of lesion in a mammogram is not only normal, benign, and malignant, but also a rank order among these categories, namely, normal < benign < cancer, as shown in Figure 1. To solve the OMIL problem, previous works have mainly focused on settling one of issues, which could be roughly divided into two problems: one is multiple instance classification, and the other is the ordinal classification problem.

Multi-instance classification is the most widely used MIL paradigm. Zhou et al. [4] proposed a MIEN-metric method that learns discriminative metrics for classifying samples from the observed classes and recognizing the samples from the novel class, which solves the emerging novel class problem concerning multi-instance learning in open and changing complex scenarios. Subsequently, in the study of multi-instance classification, Zhou et al. [5] found that samples in multi-instance learning are described by multiple instances and associated with multiple class labels. Based on a simple degeneration strategy with MIMLBOOST and MIMLSVM algorithms, the proposed multi-instance multi-label framework can deal with problems involving complex objects with multiple semantic meanings. Class imbalance is also a persistent problem in multi-instance classification tasks. Javad et al. [13] proposed a new instance reduction method that preserves the between-class distribution in the balanced data and handles minority class instance reduction in two-class imbalanced data.

The method of multi-instance learning for classification problems generally assumes that the class value is unordered. However, in many practical applications, a natural ordered relationship exists between the categories of instances. Frank et al. [14] proposed a simple method involving the incorporation of a decision tree. Accordingly, the standard classifier algorithm made full use of the ordered information between the class attributes. Different from most of the previous works that employ a loss function based on the absolute difference between the predicted and ground truth class labels, Joan et al. [15] argues that the label values in ordinal classification may be arbitrary, and proposed a network architecture that produces not a single class prediction but an ordered vector.

The recent work by Evan [12] used OMIL in the estimation of ulcerative colitis severity. However, their method transformed the multi-classification problem of OMIL into multiple binary classifiers, which fails to directly model multi-class instance bags and loses the ordered information between classes. To tackle this problem, we propose a discriminative solution to directly model ordinal multi-instance bags, which obviously learns ordered information between classes. Specifically, we exact key instances from the bag as positive instances of bags, and design ordinal minimum uncertainty loss to iteratively optimize the selected key instances from the bags. This study’s main contributions are summarized as follows:
1. We select a key instance as a positive instance of the bags and send the key instance, which has a bag label, to the network for training. The instances selected from the bag at each iteration are uncertain but incur minimal loss.
2. We employ ordinal minimum uncertainty loss to take advantage of the ordered information in classes.
3. We carry out experiments on a DDSM dataset to evaluate the OMIL method. The experimental results demonstrate that our OMIL approach achieves better performance than the existing OMIL method.

The rest of the paper is organized as follows. In Section 2, we introduce some related works about key instances, multi-instance learning, and ordinal classification. The proposed OMIL method and model architecture are proposed in Section 3. In Section 4, the processing of the DDSM dataset and experimental setup are introduced in detail. The experimental results are also presented. In Section 5, in order to verify the effectiveness of our proposed method, we conduct some ablation studies. Finally, Section 6 presents the conclusions of this paper.

![Figure 1](image_url)

**Figure 1.** We show images of three types of breast disease lesions: (a) normal, (b) benign, and (c) cancerous. Specifically, the image (bag) consists of several patches (instances). The instances marked by blue rectangles denote normal lesion, green rectangles denote benign lesion, and red rectangles denote cancer. More importantly, the lesion level of the instance in each bag cannot exceed the lesion intensity of the bag.

2. Related Works

This section defines the basis for and reviews the related works on key instances, multi-instance learning, and ordinal classification.

2.1. Key Instance

Key instances [16–18] play a key role in multi-instance learning, and their labels can trigger the label of the bag. To solve the key instance detection (KID) problem, Liu et al. [16] proposed a voting framework (VF) solution to KID, which utilizes the relationship among instances to form a citer KNN graph, and uses them to define the confidences of the votes of the training instances. However, when encountering a more complex situation, key instance detection may fail. Traditional max pooling cannot make full use of the information from input examples. Yan et al. [19] proposed a novel dynamic-pooling function for MIL that can iteratively update the instance contribution to its bag and highlights the key instance. Inspired by this work, the key instance of our model selected from each bag is constantly optimized. In order to incorporate interpretability into the MIL approach, Ilse et al. [20] proposed an “Attention-based Multiple Instance Learning” method, which pays more attention to positive instances during training. Hence, the attention weights allow us to find a key instance. Notably, this approach makes it clear how each instance contributes to the bag. Nonetheless, shin et al. [17] argued that the performance of the model with
respect to key instance detection is limited; that is, an attention-based model [20] focuses on the weight of the positive instance yet the difference between positive and negative instances in the positive bag is not obvious, which may influence performance. To improve the performance of the attention-based model in a KID task, they apply a neural network inversion with a sparseness constraint that updates the instances in a positive bag. In this way, the key instance is better highlighted by using optimized instances.

2.2. Multiple Instance Learning

Multi-instance learning [5,6,20–22] was proposed by early researchers when studying drug activity prediction. The standard assumption for multi-instance learning is that if a bag is labeled positive, there is at least one positive instance in it. Otherwise, the bag will be labeled negative. The traditional multi-instance hypothesis assumes that the class label of a bag is determined by the key instances in the bag. Multi-instance classification is the most common task in multi-instance learning. Furthermore, it has promoted the emergence of many classification algorithms, such as DD, Citation-KNN, BP-MIP, MI-SVM, mi-SVM, etc. All these MIL algorithms [21] assume that the bag is a binary classification problem, namely, positive and negative bag.

However, in general scenarios, multiple classification tasks are more common than binary classification tasks. Over the last few years, the multi-instance multi-classification task has also attracted the attention of many researchers. Different from ordinary multi-instance learning, which assumes each observation belongs to a class, a framework of multi-instance multi-label learning (MIMIL) can better describe complex objects with several classes. Zhou et al. [5] proposed the MIMLBOOST and MIMLSVM algorithms to solve problems involving complex objects associated with multiple class labels. Following this work, Pham et al. [23] extended the MIMIL problem to the setting wherein a novel class instance is present. They proposed a maximum likelihood method to optimize the model and trained an instance-level classifier for all classes as well as the novel class. Nonetheless, the ground-truth label of the sample is difficult to obtain in the real world. To alleviate this problem, Ishida et al. [24] proposed a novel setting called complementary label learning [24,25] to implement multi-class classification tasks, which only requires the provision of complementary data labels.

In this paper, we focus on the ordinal multi-instance-learning problem. The most similar work to ours is the ordinal multi-instance-learning approach proposed by Evan et al. [12]. One difference is that we directly model the bag as multi-class and select the instance with minimum loss from the bags as a positive instance of the bags to update the model parameters. In this way, each key instance we select from the bag is uncertain and optimal. Nevertheless, they transformed the multi-classification problem of bags into several binary classifiers to deal with, rather than explicitly modeling the bag.

2.3. Ordinal Classification

Ordinal classification problems [15,26–28] can be viewed as an intermediate problem between classification and regression, where the target variable is both categorical and ordinal. In general classification problems, the categorical variables are taken from a finite set, and there is no metric relationship between the categories, although they are represented numerically. Examples of categorical variables are gender, race, nationality, types of animals, etc. When there is a naturally ordered relationship between categorical variables, the ordinary classification problem transforms into an ordinal classification problem. Some common scenarios include bank credit-rating assessments, determining income level, the lesion grade of cancer, users’ evaluation of service, etc.

Gutierrez et al. [28] assessed the performance of five different methods under different representations of ordinal input variables. The results show that both the Num and Num-CDR methods perform well and the Num method directly maps each class to a consecutive natural number. Considering that the misclassification loss of ordinal classification should be different, Beckham et al. [27] proposed a simple modification of the squared error loss,
which utilizes the characteristic of the sensitivity of class ordering and guarantees that the possibility of distribution over the classes is discrete. Different from most previous works that compute the absolute difference between the predicted and ground-truth class labels to optimize the loss function, Joan et al. [15] argued that label values in ordinal classification may be arbitrary and replace a single class prediction with an ordered vector.

3. Proposed Approach

In this section, we will briefly formalize the OMIL problem and describe several crucial concepts of OMIL. In addition, we introduce a discriminative model for OMIL data wherein the instances in each bag are ordered.

3.1. Problem Formalization of OMIL

In the problem definition of OMIL, the training set is represented as: \( D = \{ (X_i, y_i) \mid i = 1, \ldots, N \} \), where \( X_i \) denotes a bag, which contains \( n_i \) instances, and each instance is described by a \( d \)-dimensional vector, i.e., \( X_i = \{ x_{i1}, x_{i2}, \ldots, x_{in_i} \} \), \( x_{ij} \in \mathbb{R}^d \). The labels of the bags are not only multi-class but also have a certain rank order [12] among classes. Simultaneously, the labeling of the instances in the bag cannot exceed the label of the bag. The formal definition is as follows: bag-level class label \( y_i \in y = \{ 0, 1, \ldots, C \} \), \( 0 < 1 < \cdots < C \), and instance-level class label \( y_{ij} \in y = \{ 0, 1, \ldots, C_1 \} \), \( 0 < 1 < \cdots < C_1 \leq y_i \). It is supposed that: (1) If a bag \( X_i \) is labeled as the \( c \)-th class \( \iff \) The instance-level label of \( X_i \) belongs to the set \( y' = \{ 0, 1, \ldots, C \} \), \( 0 < 1 < \cdots < C \). The strongest label-level of instances must be the same as the bag label, where the strongest instance-level label is \( c \)-th class. (2) A label for a bag \( X_i \) is not assigned to the \( c \)-th class. \( \iff \) None of the instances in bag \( X_i \) belong to the \( c \)-th class. We give a full description of the three important concepts of OMIL in Table 1.

| Concepts | Descriptions |
|----------|--------------|
| Bags     | The data unit of OMIL dataset that has labels. Each bag contains several ordinal instances without label information. |
| Instance | The label class of instances in a bag is ordered, and the largest category belongs to the bag category. |
| Key instance | The instance with minimum loss, which is selected from bag \( X_i \). Bag-level label is provided for the instance to participate in the model-training process. |

3.2. Model

The proposed model addresses the OMIL problem in two basic steps: (1) after the model outputs the loss of all the instances in the bag, the instance with the least amount of loss is selected as the key instance of the bag. (2) The selected key instances serve as a positive instance of the bag, assign the label of the bag, and participate in the training and optimization of the model.

Our model is presented in Figure 2. The training sample set is composed of a multi-class bag, which has a class label but the instances in it do not possess label information. The convolutional neural network [29] consists of two convolutional layers with a \( 5 \times 5 \) and \( 3 \times 3 \) filter, one pooling layer, and two fully connected layers [30]. The specific structural design of the CNN is shown in Table 2. In the model-training phase, not all instances in a bag participate in training, but the selected key instance [6,17] from the bags can be fed into the network to optimize parameters. In addition, the selected key instances from the bags are not constant, and they will further approach the true positive instances under each optimization of the network.
Table 2. Setting of the parameters of convolutional neural network (CNN) in the model structure diagram.

| Network Layer | Input | Output | Filter | Stride | Padding | Parameter |
|---------------|-------|--------|--------|--------|----------|-----------|
| Input         | $224 \times 224 \times 1$ | $224 \times 224 \times 1$ | $-$    | $-$    | $-$      | $0$       |
| Conv1         | $224 \times 224 \times 1$ | $220 \times 220 \times 10$ | $5 \times 5 \times 10$ | $0$ | $0$ | $260$ |
| Relu          | $220 \times 220 \times 10$ | $220 \times 220 \times 10$ | $1$ | $0$ | $0$ | $0$ |
| MaxPool       | $220 \times 220 \times 10$ | $110 \times 110 \times 10$ | $2 \times 2$ | $0$ | $0$ | $0$ |
| Conv2         | $110 \times 110 \times 10$ | $108 \times 108 \times 20$ | $3 \times 3 \times 20$ | $0$ | $0$ | $0$ |
| Relu          | $108 \times 108 \times 20$ | $108 \times 108 \times 20$ | $1$ | $0$ | $0$ | $0$ |
| FC1           | $108 \times 108 \times 20$ | $500$ | $1$ | $0$ | $0$ | $116,640,500$ |
| FC2           | $500$ | $1$ | $1$ | $0$ | $0$ | $501$ |

We illustrate the use of our notation, for example, of OMIL in Table 3, where $Y_b \in \text{argmax}\{0, 1, \cdots, C\}$. This model indicates that the bag label $Y_b$ is obtained from the strongest-level label in the instance labels $Y^n_b$. Hence, $Y_b$ can reveal information about the ordinal class of instances in the bag.

Table 3. Instance labels and bag labels of the proposed model.

| Bag 1 | Bag 2 | Bag 3 |
|-------|-------|-------|
| $n_1 = 1$ | $n_2 = 2$ | $n_3 = 3$ |
| $y_{11} = \text{normal}$ | $y_{21} = \text{normal}$ | $y_{31} = \text{normal}$ |
| $y_{12} = \text{normal}$ | $y_{22} = \text{normal}$ | $y_{32} = \text{benign}$ |
| ... | ... | ... |
| $y_{1n} = \text{normal}$ | $y_{2n} = \text{benign}$ | $y_{3n} = \text{cancer}$ |
| $Y_1 = \{\text{normal}\}$ | $Y_2 = \{\text{normal, benign}\}$ | $Y_3 = \{\text{normal, benign, cancer}\}$ |

Figure 2. Overall model structure of OOMIL. Mammograms are split into patches, and then fed to the model to select the minimum-loss patch. The key instance-assigned bag labels participate in model training.

4. Experiments

In this section, we evaluate the performance of the proposed novel OOMIL approach against the original OMIL approach, and the results of the comparison when applied to a DDSM dataset are shown in Table 4. Further on, to increase the interpretability of the model, we visualize the process of the model’s selection of key instances from a bag, which is displayed in Figure 3.
In this section, we evaluate the performance of the proposed novel OOMIL approach against the original OMIL approach, and the results of the comparison when applied to a DDSM dataset are shown in Table 4. Further on, to increase the interpretability of the model, we visualize the process of the model’s selection of key instances from a bag, which is displayed in Figure 3.

Figure 3. When the key instance $K$ value is 4, the number of positive instances $S$ in the key instances selected by the model and the model training loss change as the epoch $E$ increases. $E$, $K$, and $S$ represent epoch times, key instance, and positive instance in key instance, respectively. The key instances marked with a blue rectangle denote false instances, whereas the red rectangles represent true instances.
Table 4. Test accuracy comparison results for OMIL and OOMIL (our) method.

| Dataset | OMIL     | OOMIL (Our Method) |
|---------|----------|---------------------|
| DDSM    | 49.882   | 52.021              |

4.1. Datasets

DDSM [31] is a widely used mammography dataset in computer-aided medical diagnosis. DDSM includes four types of data, namely, normal, benign, benign-without-callback, and cancer, which comprise 10,420 images from 2605 breast cancer cases. For benign and cancer cases, only the images with the lesion area marked by a physician are selected. Hence, our dataset employs 1700 images from normal, benign, and cancer cases. We selected 1360 images from each type of case to form the training set and 340 images to form the test set.

Some processing operations must be performed on the training set. We first divide each image into grids [32–34] with an aspect ratio of 14:7 to obtain 98 patches so that each bag (image) consists of 98 instances (patches). Since some instances come from the noisy region of the image, we use a threshold-processing operation [35,36] to further filter the instances in the bag. To ensure a balanced number of instances per bag, we again expand the dataset with horizontal and vertical flipping as well as rotating image enhancement techniques [37,38], and resize each instance to 224 × 224. Finally, our training set consists of 4080 bags (1360 per lesion type), each containing 70 instances with a 224 × 224 size. To verify the effectiveness of the instance-level classifier trained by our proposed method, we crop the region of interest from each image in the test set as a test example, which has a specific class label.

4.2. Experimental Setup

In our experiments, our network is composed of two convolution layers with a 5 × 5 and 3 × 3 filters, a pooling layer with a 2 × 2 kernel, and two fully connected layers [30] with 500 neurons and 1 neuron, respectively. The weights and biases are initialized to be 0. In addition, the k value of the selected key instance is assigned to 4. We employ the approach developed by Adam [39] to optimize our network. Detailed hyper-parameters such as the learning rate, weight decay, eps, and betas are set to 0.0001, 0.0005, and 1 × 10⁻⁶, (0.9, 0.99). We choose pytorch as the deep-learning framework [40] and write the code in python. The experiments are run on a PC with AMD EPYC-7302 CPU and 64 GB RAM.

4.3. Empirical Results of OMIL

Here, we present the comparison results for our novel OOMIL method and existing OMIL method. The experimental results in Table 4 show that our method performs better than the OMIL when applied to the DDSM dataset. In order to intuitively show the effectiveness of our proposed method, we visualize the process of the model selecting the key instances from a bag. The process is shown in Figure 3. The letter E represents the training epoch of the model, K represents the number of key instances to pick from the bag, and S denotes the number of key instances that represent positive instances. Instances with a blue rectangle represent the key instances that are not selected correctly. On the contrary, the red rectangles represent key instances that are selected correctly. As is vividly shown in Figure 3, when the training epoch increases, the number of real key instances that are selected out also increases. Thus, the effectiveness of our proposed method has been proven.

5. Ablation Study and Discussion

In this section, we conduct some ablation studies regarding the effect of the k value of the key instance on each class prediction and different loss functions across the DDSM dataset. In addition, the evaluation metrics of the model are presented in Table 5.
Table 5. Different performance measurements of our model.

| Proposed Method | Sensitivity (%) | Specificity (%) | Precision (%) | F1 (%) |
|-----------------|-----------------|-----------------|---------------|--------|
| OOMIL           | 61.471          | 47.206          | 57.895        | 59.629 |

Each class accuracy on different k values of the key instances. As mentioned in Section 3, a key instance \([16,18]\) is an instance with minimum loss, which is closest to the positive instance in the bag. The accuracy of the key instance’s selection further affects the effectiveness of the model’s ability to learn. To reduce the probability of the erroneous selection of key instances, we select the first k instances with minimum loss as the key instances of the bag. Hence, we carry out comparison experiments to study the influence of different k values. In Table 6, we show each class result of the different k values on the DDSM dataset. In addition, we can see that as the number of key instances selected increases, each category has a more even prediction probability instead of being inclined towards selecting a certain class. That is, the model performs well overall.

Table 6. Class accuracy comparison of number set K for key instances in the model. K takes values from 1 to 5, which denote different values of selected key instances.

| Class | K              |
|-------|----------------|
|       | 1 2 3 4 5      |
| 0     | 20.294 43.529 57.059 52.059 71.765 |
| 1     | 26.176 32.941 22.059 42.353 35.882 |
| 2     | 85.294 54.706 60.000 49.118 22.353 |

Different loss functions’ accuracies on different k values of key instance. Through previous MIL tasks \([4,23]\), cross entropy loss \([41]\) has proven to be one of the most commonly used loss functions in MIL methods. Instead, our OMIL uses the ordinal minimum uncertainty loss during training. To discuss the influence of different loss functions, we conduct ample comparative experiments on cross entropy loss and minimum uncertainty ordinal loss. In order to apply the Cross Entropy loss to our OMIL, we change the network architecture by replacing the last fully connected layer, which has one neuron, with a fully connected layer, which has three neurons and employs a SoftMax function. Then, combining the bag label with the instance probability, we can calculate the Cross Entropy loss. Finally, we can obtain the loss of each instance from the bag and obtain the instance with minimum loss. Table 7 indicates that, with respect to the DDSM dataset, the results based on the minimum uncertainty ordinal loss are better than cross entropy loss. This proves the effectiveness of the minimum uncertainty ordinal loss to our OMIL.

Table 7. Average test accuracy comparison of different loss functions on number set K for key instances. K ranges from 1 to 5.

| Loss                      | K              |
|---------------------------|----------------|
|                           | 1 2 3 4 5      |
| cross entropy (OOMIL)     | 47.255 47.059 49.706 51.176 47.667 |
| min-uncertainty ordinal (OOMIL) | 48.529 47.451 49.679 52.021 47.983 |
| cross entropy (OMIL)      | 46.972 47.183 48.127 48.743 46.430 |
| min-uncertainty ordinal (OMIL) | 47.943 47.372 49.232 49.882 46.846 |

The experimental results in Table 7 show that the accuracy of the model does not always rise with the number of key instances (K). A possible reason is that the sample
imbalance affects the improvement of the model’s performance. In our training bags, the number of normal instances is much greater than that of benign and cancer instances, which is due to the limitations of the mammogram itself. The lesion regions in the mammogram may account for only 4% of the whole image. The model performs well and the prediction accuracy for each class is relatively average when the value of K is 4. The model tends to predict class 0 in the case of K being 5. The reason is that it is difficult to pick the correct instances in the early stage of model training, and the model’s fault tolerance rate is low if the number of key instances is too small. However, when the number of key instances is too large, false key instances are far larger in number than true key instances, which also leads to the degradation of the model’s performance.

Currently, most of the existing work on breast cancer classification addresses the binary classification problem, that is, judging whether the ROI (mass or calcification) is benign or malignant. Elmoufidi [1] proposed a framework comprising two steps of ROI segmentation and feature extraction from ROI for classification. The experimental results regarding the method’s sensitivity, specificity, and accuracy are 94.46%, 94.40%, and 94.43%. Fahssi et al. [2] presented a novel CAD system for mammography diagnosis. They first partition the mammogram into regions and detect the ROI, and then use MIL algorithms to identify malignant regions in order to assign the label to the whole mammogram. This achieves 90.84% sensitivity, 90.17% specificity, and 90.33% accuracy. However, in practical situations, normal cases are generally more frequent than abnormal cases in the diagnosis of breast disease. There is no doubt that screening for abnormal mammograms from large numbers of normal mammograms is energy-consuming work for physicians. Accordingly, the classifier we trained was three-class, classifying mammograms as normal, benign, or malignant. The three-class task is more challenging than ordinary binary classification, as the instance features of normal and benign categories are very similar, especially with respect to benign lesions in early breast cancer. Lamard et al. [3] carried out experiments on binary classification, which achieved an accuracy of 91.1%, while the accuracy of the three-class classification was only 62.1%, a drop of almost 30%, which strongly suggests that three-class classification tasks are more difficult than binary classification tasks. Although our experiment achieved only 52.021% accuracy, to the best of our knowledge, we are the first to propose a novel OMIL method to directly model multi-category bags rather than transforming them into multiple binary classifiers for multi-categories. During the experiment, we found that the accuracy of malignant classification is generally higher than normal and benign, while the accuracy of normal and benign categories is similar. How to more effectively classify normal and benign mammograms and improve the accuracy of model prediction is the key point to consider in our next experiment.

6. Conclusions

In this paper, we addressed the ordinal multi-instance-learning problem in breast cancer diagnosis using mammograms. Compared with the binary classification of mammograms, the three-classification task provides superior support for the screening of early breast cancer, as the features of normal tissues and benign lesion areas are very similar. During the experiment, we found that the predictive accuracy with respect to the malignant categories is generally higher than that of the normal and benign categories. Simultaneously, the accuracy gap between the normal and benign categories is small. This also shows that benign lesions can easily be predicted as normal categories, which is very influential in preventing further deterioration in early breast cancer. Hence, we propose a method that directly models the ordinal class information for bags and the instances in the bags. Moreover, to ensure that the key instance selected in the bag is closer to the real positive instances, we employed minimum uncertainty ordinal loss to iteratively optimize the selected key instances from the bags. To increase the interpretability of the model, the simple grid segmentation method acted as the generator of bags, which allowed for the convenient recording of the specific location of instances in the bag to visualize the process of the model’s instance selection. Other more elaborate bag generation methods such as
SBN, ImaBag, BlobBag, etc., may improve the performance of the model. The problem of sample imbalance is also not addressed by our current work. The number of normal samples is much higher than benign and cancer samples, leading to the higher prediction accuracy of the model for normal instances. A model with strong generalizability is trained by introducing cost-sensitive methods that impose different penalties for different numbers of samples, which is the focus of our future work. Overall, our method breaks down the barrier that medical images are difficult to label and supply help for breast cancer diagnosis in the future.

Author Contributions: Conception and design of the study, X.X. and Q.G.; collection of the data, Q.G. and D.L.; analysis of the data and contribution to the writing of the manuscript, X.X., Q.G., Z.L. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 61976217), the Fundamental Research Funds of Central Universities (No. 2019XKQYMS87), and the Science and Technology Planning Project of Xuzhou (No. KC21193).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elmoufidi, A. Deep Multiple Instance Learning for Automatic Breast Cancer Assessment Using Digital Mammography. *IEEE Trans.Instrum. Meas.* 2022, 71, 4503813. [CrossRef]

2. Elmoufidi, A.; El Fahssi, K.; Jai-andaloussi, S.; Sekkaki, A.; Govenole, Q.; Lamard, M. Anomaly Classification in Digital Mammography Based on Multiple-Instance Learning. *IET Image Process.* 2018, 12, 320–328. [CrossRef]

3. Sánchez de la Rosa, R.; Lamard, M.; Cazuguel, G.; Coatrieux, G.; Cozic, M.; Quellec, G. Multiple-Instance Learning for Breast Cancer Detection in Mammograms. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 7055–7058.

4. Wei, X.-S.; Ye, H.-J.; Mu, X.; Wu, J.; Shen, C.; Zhou, Z.-H. Multiple Instance Learning with Emerging Novel Class. *IEEE Trans. Knowl. Data Eng.* 2019, 33, 2109–2120. [CrossRef]

5. Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; Li, Y.-F. Multi-Instance Multi-Label Learning. *Artif. Intell.* 2012, 176, 2291–2320. [CrossRef]

6. Zhou, Z.-H. Multi-Instance Learning: A Survey; Technical Report; Department of Computer Science & Technology, Nanjing University: Nanjing, China, 2004.

7. Bilen, H.; Vedaldi, A. Weakly Supervised Deep Detection Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2846–2854. [CrossRef]

8. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Miraflores, V.; Erkens, J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nat. Med.* 2019, 25, 1301–1309. [CrossRef] [PubMed]

9. Kim, J.; Kim, H.J.; Kim, C.; Lee, J.H.; Kim, K.W.; Park, Y.M.; Kim, H.W.; Ki, S.Y.; Kim, Y.M.; Kim, W.H. Weakly-Supervised Deep Learning for Ultrasound Diagnosis of Breast Cancer. *Sci. Rep.* 2021, 11, 24382. [CrossRef] [PubMed]

10. Zhou, Z.-H. A Brief Introduction to Weakly Supervised Learning. *Natl. Sci. Rev.* 2018, 5, 44–53. [CrossRef]

11. Alpaydın, E.; Chelpygina, V.; Loog, M.; Tax, D.M.J. Single- vs. Multiple-Instance Classification. *Pattern Recognit.* 2015, 48, 2831–2838. [CrossRef]

12. Schwab, E.; Cula, G.O.; Standish, K.; Yip, S.S.F.; Stojimirovic, A.; Ghanem, L.; Chehoud, C. Automatic Estimation of Ulcerative Colitis Severity from Endoscopy Videos Using Ordinal Multi-Instance Learning. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* 2022, 10, 425–433. [CrossRef]

13. Hamidzadeh, J.; Kashefi, N.; Moradi, M. Combined Weighted Multi-Objective Optimizer for Instance Reduction in Two-Class Imbalanced Data Problem. *Eng. Appl. Artif. Intell.* 2020, 90, 103500. [CrossRef]

14. Frank, E.; Hall, M. A Simple Approach to Ordinal Classification. In *Machine Learning: ECML 2001*; De Raedt, L., Flach, P., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2001; Volume 2167, pp. 145–156.

15. Serrat, J.; Ruiz, I. Rank-Based Ordinal Classification. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8069–8076.

16. Liu, G.; Wu, J.; Zhou, Z.-H. Key Instance Detection in Multi-Instance Learning. In Proceedings of the Asian Conference on Machine Learning, PMLR, Singapore, 11–17 November 2012; pp. 253–268.

17. Shin, B.; Cho, J.; Yu, H.; Choi, S. Sparse Network Inversion for Key Instance Detection in Multiple Instance Learning. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4083–4090.
18. Zhang, Y.-L.; Zhou, Z.-H. Multi-Instance Learning with Key Instance Shift. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 19–25 August 2017; pp. 3441–3447.

19. Yan, Y.; Wang, X.; Gao, X.; Fang, J.; Liu, W.; Huang, J. Deep Multi-Instance Learning with Dynamic Pooling. In Proceedings of the 10th Asian Conference on Machine Learning, PMLR, Beijing, China, 14–16 November 2018; pp. 662–677.

20. Ilse, M.; Tomczak, J.; Welling, M. Attention-Based Deep Multiple Instance Learning. In Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2127–2136.

21. Astorino, A.; Fuduli, A.; Gaudio, M.; Vocaturo, E. Multiple Instance Learning Algorithm for Medical Image Classification. *SEBD 2019*, 2400, 1–8.

22. Zhou, Z.-H.; Zhang, M.-L. Neural Networks for Multi-Instance Learning. In Proceedings of the International Conference on Intelligent Information Technology, Beijing, China, 22–25 September 2002; pp. 455–459.

23. Pham, A.T.; Raich, R.; Fern, X.Z.; Arriaga, J.P. Multi-Instance Multi-Label Learning in the Presence of Novel Class Instances. In Proceedings of the International Conference on Machine Learning(ICML), Lille, France, 7–9 July 2015; pp. 2427–2435.

24. Ishida, T.; Niu, G.; Hu, W.; Sugiyama, M. Learning from Complementary Labels. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.

25. Gao, Y.; Zhang, M.-L. Discriminative Complementary-Label Learning with Weighted Loss. In Proceedings of the International Conference on Machine Learning(ICML). PMLR, Virtual, 18–24 July 2021; pp. 3587–3597.

26. Albuquerque, T.; Cruz, R.; Cardoso, J.S. Ordinal Losses for Classification of Cervical Cancer Risk. *Peerf Comput. Sci. 2021*, 7, e457. [CrossRef] [PubMed]

27. Beckham, C.; Pal, C. A Simple Squared-Error Reformulation for Ordinal Classification 2017. *arXiv 2016*, arXiv:1612.00775.

28. Gutiérrez, P.A.; Pérez-Ortiz, M.; Sánchez-Monedero, J.; Hervás-Martínez, C. Representing Ordinal Input Variables in the Context of Ordinal Classification. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 2174–2181.

29. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE 1998*, 86, 2278–2324. [CrossRef]

30. Jegou, S.; Drozdzal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.

31. Lekamlage, C.D.; Afzal, F.; Westerberg, E.; Chaddad, A. Mini-DDSM: Mammography-Based Automatic Age Estimation. In Proceedings of the 2020 3rd International Conference on Digital Medicine and Image Processing, ACM, Kyoto Japan, 6–9 November 2020; pp. 1–6.

32. Bhagat, P.K.; Choudhary, P.; Singh, K.M. Two Efficient Image Bag Generators for Multi-Instance Multi-Label Learning. In Proceedings of the Computer Vision and Image Processing, Jaipur, India, 27–29 September 2019; Nain, N., Vipparthi, S.K., Raman, B., Eds.; Springer: Singapore, 2020; pp. 407–418.

33. Wei, X.-S.; Zhou, Z.-H. An Empirical Study on Image Bag Generators for Multi-Instance Learning. *Mach. Learn. 2016*, 105, 155–198. [CrossRef]

34. Zhou, Z.-H.; Zhang, M.-L.; Chen, K.-J. A Novel Bag Generator for Image Database Retrieval with Multi-Instance Learning Techniques. In Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, Sacramento, CA, USA, 3–5 November 2003; pp. 565–569.

35. Jyothi, S.; Bhargavi, K. A Survey on Threshold Based Segmentation Technique in Image Processing. 26 K Bhargavi Jyothi. *Int. J. Innov. Res. Dev. 2014*, 3, 234–239.

36. Middleton, D. Channel Modeling and Threshold Signal Processing in Underwater Acoustics: An Analytical Overview. *IEEE J. Ocean. Eng. 1987*, 12, 4–28. [CrossRef]

37. Hong, L.; Wan, Y.; Jain, A. Fingerprint Image Enhancement: Algorithm and Performance Evaluation. *IEEE Trans. Pattern Anal. Mach. Intell. 1998*, 20, 777–789. [CrossRef]

38. Maini, R.; Aggarwal, H. A Comprehensive Review of Image Enhancement Techniques 2010. *arXiv 2010*, arXiv:1003.4053.

39. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization 2017. *arXiv 2014*, arXiv:1412.6980.

40. Jaafar, B.; Mahersia, H.; Lachiri, Z. A Survey on Deep Learning Techniques Used for Breast Cancer Detection. In Proceedings of the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sfax, Tunisia, 2–5 September 2020; pp. 1–6.

41. Zhang, Z.; Sabuncu, M. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Curran Associates, Inc.: New York, NY, USA, 2018; Volume 31.